

Mikhail Gorbachev's tightest corner

This week and those immediately ahead should see at least some decisions made, or alternatively taken by default, about the future of the Soviet Union and its leader, Mr Mikhail Gorbachev.

LAST week's first visit to Japan by a Soviet head of state was predictably a disappointment, given that the Soviet Union had said in advance that it would not concede Japan's case for the restoration of the disputed Kurile islands, annexed in 1945, but Mr Mikhail Gorbachev can be pleased about at least one feature of the visit: his enemies at home did not seize on his absence as an opportunity to depose him from one or other of his powerful positions as president of the Congress of Deputies and secretary general of the Soviet Communist Party. All too often, people in his position have found themselves deposed while glad-handing in foreign capitals. But Gorbachev cannot carry on much longer unless he manages to break out of the pattern of inanition into which he has slumped in the past few months.

The most tangible symptoms of his difficulty are economic. Last week, the government allowed that industrial production in the Soviet Union has fallen in a year by 8 per cent, comparable with the extent to which Western economies were depressed during the 1930s. At the same time, exports have fallen by two-thirds and the government is quickly running through the stock of freshly printed roubles it had hoped would last for the whole year. The prospect of an economic catastrophe on a scale even greater than the appalling situation of the past year stares the Soviet Union in the face.

Opportunity

What can be done? On this narrow front, Gorbachev's weakness is that he has let slip past opportunities to put the Soviet economy to rights. The 'price reforms' introduced last month, which are not really reforms but merely new and higher prices centrally decreed, were first planned in 1987. Then, they might have worked. Now, after four years of needless delay, they seem merely to have precipitated a series of strikes that will further disrupt industrial production.

More recently, a year ago, Gorbachev blew hot and cold (in that order) about the plans for the more radical reform of the economic structure being canvassed by people such as Academicians S. S. Shatalin and A. G. Aganbegyan. Having encouraged the idea, fostered by Mr Boris Yeltsin, that the rudiments of a market economy should be introduced as the first step in a whirlwind upheaval lasting just 500 days, Gorbachev first hijacked the idea and then abandoned it. Not surprisingly, that proved a recipe for losing friends and helpers. Now, towards the end of a miserable winter, even the struggling free-market cooperatives have been further constrained by decree from above. Even if Gorbachev now

wished to resuscitate those lost opportunities, it is doubtful whether the Communist Party and the *soyuz* faction in the Congress would let him.

Separation

At this late stage, in any case, an exclusively economic reform, however radical, will not put the Soviet Union and its several increasingly independent parts to rights. Assertions of separation by the three Baltic states and Georgia remain in a kind of limbo, while the relationship between the old command centre — the Central Committee and the Soviet government — with the republics and especially Yeltsin's Russian Federation, are similarly unresolved. Meanwhile, the fabric of an ingenious society hardy enough to have survived 70 dreadful years is being quickly eroded by people's growing perception that they must fend for themselves. In all the circumstances, the surprise is not that Soviet science seems to have benefited little from *glasnost* (which persists), but that some heroic people are still to be found working at the bench. For how much longer?

The coming weeks should settle some of these undecided issues. For Gorbachev, and for the Soviet Union as a whole, this Wednesday's plenum of the Communist Party will be crucial. From the outset, Gorbachev's ambivalence about the party's role in the Soviet scheme of things has been a source of confusion. His chief constitutional innovation has been to create the Congress of Peoples' Deputies, in principle at least an elected parliament, but the party retains the right to nominate government ministers (who are then approved, or otherwise, by the Congress).

Four years ago, Gorbachev might have risked resigning his post as general secretary, relying on the compliance of a discredited organization to yield governments that would have been acceptable or, even better, transferring that crucial function from the party to the Congress. His failure to grasp that nettle probably explains why he has become the prisoner of party opinion on centralism, the hankering after the concept of a planned imperial state to which two tiny islands off the coast of Japan are indispensable. Now, resignation would risk confrontation between the party and centrifugal nationalism on the one hand and democratic liberalism on the other. That is one of several ways in which civil war might start. The best hope is that Gorbachev stays on, and is allowed to do so, at least for the time being.

But that can be only half the battle. There also needs to be a plan to make the Soviet Union function. Those who at the



outset welcomed Gorbachev's *glasnost* and his plans for *perestroika*, as if dismayed at the degree to which he has been weakened by the ferocity of their protests at what seemed to be backsliding, are now offering the olive branch of some kind of coalition government. If it could be realized, it might just work. But the chances are only small, while those of breathing new life into one of the discarded economic plans have been much reduced by the widespread sense of having been monkeyed with that Soviet people now widely share. □

Soft on software

Europe, which plans to protect software against piracy, should now give some thought to those who buy it.

EARLIER this month, the European Commission did the decent thing by the international software industry by requiring member states to adopt uniform copyright laws on commercial software. Both equity and commercial prudence require nothing less. To copy a piece of commercial software illicitly is theft of a literal kind, where the victims are those who have invested skill, time and money in the development of a useful set of computer instructions. The uniform extension of copyright legislation to software is also commercially prudent; otherwise, the supply might dry up. Software manufacturers say they are pleased with the development, but also (rightly) that enforcement is not child's play.

So far, so good. But before tears are shed for software manufacturers, some thought should be given to their licit customers, and to the terms on which software packages are sold. The use of the term copyright suggests an analogy with books, but the analogy is inexact in at least two ways. First, a book (at least from a reputable publisher) is a finished product that is unlikely to be substantially revised for at least some time; software packages, on the other hand, are published almost breathlessly, with 'version numbers' growing steadily in each of their two significant digits. Second, books enjoy well-understood economies of scale, with the result that publishers price their products so as to make money if their expectations are requited, but otherwise lose. By contrast, the intrinsic value of a software package is negligible compared with the selling price.

These circumstances put software purchasers at a disadvantage compared with mere book-buyers. For one thing, having bought a software package advertised as serving some stated purpose, they must then pay extra to upgrade to a new version that will actually do the job. Second, they derive no benefit from economies of scale. Whereas a successful book will probably appear eventually more cheaply as a paperback, a successful and widely used software package will keep its original high price until some other manufacturer is able to achieve the same results with a separately written set of codes. And as things are, it is not possible for people to buy only part of a software package, perhaps to carry out only one of many functions. The European Commission, which is often fierce in its execution of anticompetition policy, should now look at the other side of the software coin. □

Cholera marches on

The outbreak of cholera in northwestern South America could cast a long shadow on the world.

CHOLERA is always with us. That is literally the case. The causative organism, *Vibrio cholera*, is widely spread among the human population, but the enterotoxin that the organism secretes is likely to cause intestinal disease only when there are as many as 100 million organisms in the gut. The consequence is that cholera organisms can be widely spread yet cryptic in a population until some external agency, usually a deterioration of hygiene, causes them to multiply. In the early decades of the nineteenth century, the disease was commonplace in major European cities. Even now, it is endemic in cities such as Calcutta, where sporadic cases repeatedly crop up. But what is to be made of the outbreak of cholera since January in the South American states of Chile, Peru, Ecuador, Colombia and, now, some parts of Brazil?

More than 150,000 people have been affected in Peru alone, of whom more than 1,000 have died. It seems to have been recognized by the governments of the region, which met last week in Bolivia, that the necessary remedies are those well understood for more than a century — supplies of clean water, foodstuffs similarly free from contamination and better sanitation. But this simple recipe is more easily stated than made reality. Otherwise, the present outbreak, which is but the latest manifestation of a pandemic that began in Indonesia in 1961 and which caused 150,000 cases in Africa a decade later, might have been avoided.

Dr John Wood, the London physician, made himself a legend when, in 1854, he removed the pump-handle from a polluted communal water supply in the City of London's Broad Street, bringing a cholera outbreak quickly to an end, but plainly there can be no such simple solution in the vast region already affected in South America. Public education in the principles of personal hygiene is the only ready expedient. The nightmare — made explicit at the weekend by an official of the Panamanian Health Organization — is that the outbreak might spread to the great cities of the Amazon basin and, from there, to Brazil's shameful shanty towns around São Paulo and Rio de Janeiro.

There is an even more daunting prospect. Iranian health officials are already alarmed by cases of cholera in southern Iraq, among the people recently afflicted by the Gulf War and its aftermath. (Whether cases are officially reported depends on the willingness of governments to take such invidious action.)

Already fragile public health arrangements have deteriorated and, worse, the tragic displacement of people from their homes has made them dependent on makeshift and insecure sanitary systems. In general, cholera is an attribute of poverty, and of the lack of public investment in water supplies and sanitation, but social upheavals like those to be expected in the Middle East in the next several years are also precipitating causes. We shall hear a lot more of cholera before much more time passes. □

Compromise averts Framework fiasco

■ Five research programmes back on track

■ A continuing power struggle over policy

London

THE warring institutions of the European Communities (EC) have managed to patch together a compromise that should prevent the withdrawal, announced by the European Commission last month, of five planned research programmes (see *Nature* 350, 177; 21 March 1991).

The withdrawal of the research programmes, the only five among the 15 of the EC's third five-year Framework research budget that are nearing final approval, would have delayed any spending of the budget of about £4,000 million (5,700 million European Currency Units) until at least late 1992. The third Framework was intended originally to run from 1990 to 1994.

Although the immediate crisis has been resolved, the third Framework remains badly behind schedule, and further political upheaval is possible. Most of the arguments over EC research stem from the perennial struggle for power between the ministers from EC member states, the European Parliament and the Commission, which is the EC's civil service. With the EC's constitution due for redrafting, these battles are likely to intensify, as the member states debate plans for greater political unity.

The recent dispute over the research budget began when the Commission, backed by the European Parliament, said it was withdrawing the five programmes because the research ministers from the EC member states had altered the programmes from the Commission's original proposals, and had ignored amendments suggested by the Parliament. But the move was seen by many as a calculated act of defiance by the Commission and the Parliament, using the EC's research programme as a political battleground to press their case for increased power over EC policy — which is now dominated by the decisions taken at the occasional meetings of ministers from the EC states.

Jacques Delors and Enrique Baron Crespo, presidents of the Commission and the Parliament respectively, and Luxembourg's research minister Rene Steichen last week signed an agreement resolving the differences over the five research programme proposals, which is expected to be endorsed this week at meeting of EC research ministers.

Both sides claim to have given away little in the compromise. But some observers in Brussels believe that Delors, having made a political point, has now decided to leave the EC's research programmes alone.

The Commission and Parliament had dis-

agreed with the ministers over a number of points concerning the management of the five research programmes. The main concession made by the ministers is to weaken some of the management committees that will oversee the programmes in the third Framework (not just the five disputed programmes) — so allowing the Commission more freedom to control research programmes.

The five threatened programmes, in environmental research, marine science and technology, life sciences, technologies for developing countries, and two programmes in telecommunications, should now be finally adopted in June, and the first grants may be awarded early next year. But the delays introduced by a series of skirmishes between the member states' ministers and the Commission mean that the grant spending under the rest of the programmes within the third Framework is unlikely to begin before mid-1992. This makes a mockery of the original plan to review the progress of each of the research programmes during 1992, halfway through the third Framework.

Nevertheless, Gordon Adam, vice chairman of the Parliament's Energy, Research and Technology Committee, believes the crisis of the past month may have had one positive effect. To forge the compromise, research ministers and members of the European Parliament have been forced to communicate with one another. Adam hopes the precedent will lead to a more open dialogue between the two groups.

But a further row over EC research is likely. A Commission proposal to have future Framework budgets and the details of their component programmes approved in a single step (see *Nature* 350, 101; 14 March 1991) has been included in a draft revision of the EC's constitution, now being circulated around the EC. The draft has been produced by the intergovernmental conference that is considering plans for greater EC political unity. At present, the proposed outlines of individual research programmes are subject to scrutiny by research ministers and the Parliament after the overall Framework budget is agreed. If each Framework was approved in a single step, the Commission would have far more influence over EC research than is now the case.

Given the delays that have plagued the approval of the programmes that make up the third Framework, some EC states are expected to support the move to streamline the process. But British officials, fearing any extension of the Commission's power, will oppose the plan.

Peter Aldhous

EUROPEAN RESEARCH

Setting up shop in Brussels

London

RESEARCH funding agencies from the European Communities (EC) member states are stepping up their presence in Brussels, anticipating an increase in the proportion of research spending by European governments coming through the EC.

The five British research councils last week merged their Brussels office with that of the British Council, which has been speaking with the EC on behalf of a number of British universities. The new UK Research and Higher Education European Office aims to ensure that UK scientists are able to compete effectively for EC research grants.

British scientists — aided by the research councils' presence in Brussels — currently earn more in EC research grants than the UK government pays into the various EC research programmes. The French and Germans have so far been less successful in winning EC funding, but now seem to be following the British example. French research funding agencies opened a Brussels office last month, and the Germans are soon to follow suit. P.A.

RESEARCH ETHICS

Nuffield takes the lead on bioethics

London

THE Nuffield Foundation, a large British research charity, will set up the first UK committee to consider the broad sweep of ethical issues thrown up by advances in biological research. The Nuffield Council on Bioethics will be chaired by Sir Patrick Nairne, formerly a senior civil servant at the UK Department of Health and Social Security, and will have a budget of some £150,000 a year.

David Shapiro, from the Nuffield Foundation's bioethics initiative, expects that the Nuffield Council will concentrate initially on the ethical questions surrounding new techniques in molecular biology, including genetic screening, and on publishing reports of its work. A majority of the council's members will be neither professional scientists nor medical doctors, Shapiro says.

The Nuffield Council breaks with the British custom of setting up separate *ad hoc* committees to tackle different ethical issues — for example, the Interim Licensing Authority for human embryo research, set up by the Medical Research Council and the Royal College of Obstetricians and Gynaecologists.

Perhaps the closest parallel for the new body is the French National Consultative Ethics Committee for the Life Sciences and Health, set up in 1983 — except that the French body is supported by public funds, rather than a charitable foundation. P.A.

Cuts coming on overheads

Washington

A CAP on the indirect costs of research, only a rumour two weeks ago, now seems nearly certain as US government officials and congressional leaders rush to respond to the overcharging scandal that is striking a growing number of US universities.

At a series of congressional hearings last week, Bush Administration officials launched a preemptive strike against feared congressional initiatives by proposing their own across-the-board caps on indirect costs. A task force on the subject chaired by Bernardine Healy, the new director of the National Institutes of Health (NIH), is seriously considering limits on total costs as well as restrictions on the portion of those costs that cover administrative expenses.

Richard Kusserow, inspector general of the Department of Health and Human Service (HHS) and a member of Healy's task force, testified that capping indirect costs at 50 per cent of direct research costs would save some \$125 million each year, which could fund about 1,000 new research grants at NIH. Although 50 per cent is about the current average for indirect costs, declaring it an absolute ceiling would force universities above the average to lower their rates, resulting in an overall reduction in cost to the federal government — a reduction that universities will have to make up from their own resources.

Other proposals also remain under active consideration by the Healy panel and the White House Office of Management and Budget, although they are seen as less likely than the option of capping indirect costs. These alternatives include 'block grants' to universities, in which researchers and administrators would decide between themselves how much should go to direct and indirect costs, as well as grants awarded directly to researchers, who would then negotiate with their administrators what proportion would go to indirect costs.

Five years ago, the federal government proposed and nearly succeeded in winning a cap on the administrative portion of indirect costs, following the recommendations of a report by D. Allan Bromley (now director of the White House Office of Science and Technology Policy) and industrialist David Packard. At the time, such a cap would have saved \$85 million per year — perhaps half of what a similar measure might save today in the larger NIH budget.

Last week, that proposal was resurrected when Kevin Moley of HHS suggested at least a 26 per cent cap on administrative indirect costs, which currently make up about half of total indirect costs. Given the accounting reforms since 1986, a ceiling on administrative costs might even be profitably lowered further, Moley told a congressional committee. HHS is also said to be considering a 20 per cent level.

Although that sort of proposal had full federal support in 1986, Congress, under university pressure, put language in its appropriations bill to prohibit it. This time, however, in the midst of an indirect-cost scandal, Congress seems unlikely to offer much opposition. Indeed, there is an unmistakable groundswell of determined — if somewhat undirected — congressional enthusiasm for cost cuts.

Science supporters fear that congressional appropriations committees, faced with deficit troubles and a host of competing financial demands, may simply reduce the NIH budget in anticipation of the savings from whatever cost-control plan is finally adopted. That money would presumably go to fill needs elsewhere in the US budget.

Science supporters in Congress are

anxious to convince colleagues of the dangers of hasty cuts. "Most members of Congress", says George Brown (Democrat, California), chairman of the House Science Committee and strong research supporter, "don't understand the issue and don't understand why the money [saved by capping indirect costs] should go back into science. As the possibility of a cap dawns, some of us will be trying to educate them."

Meanwhile, university administrators, fearing that the issue will soon be out of their hands, are trying to be seen as responsible. The Association of American Universities last week told congressional leaders that it has asked a panel of university presidents to visit Washington and meet federal officials "to correct errors which call into question our commitment to contain and monitor prudently and responsibly the costs of conducting sponsored research."

Christopher Anderson

ANIMAL RESEARCH

Court puts rules in question

Washington

A US COURT victory earlier this month by an animal-welfare organization has opened the door to a legal battle over federal restrictions on the care of rats, mice and birds in research, as well as a possible attack on the rules covering primates, dogs and cats.

The suit, filed in June last year by the Animal Legal Defense Fund, sought to force the US Department of Agriculture (USDA) to include rats, mice and birds in its animal care and housing regulations, which govern all US laboratory animals.

A clause in the 1971 congressional act that compels USDA to protect the welfare of laboratory animals gives the administrator of the agency the authority to determine what is and is not to be defined as a research animal. To simplify its enforcement, USDA has chosen not to define rats, mice and birds as 'animals' for regulatory purposes. The agency argues that including them would greatly increase its administrative workload without greatly improving their welfare, as most laboratory rat and mouse enclosures are already constructed in compliance with the accepted guidelines of the National In-

stitutes of Health.

In its decision earlier this month, the US District Court for the District of Columbia ruled that the animal defence group has the right to sue because it had been prevented from distributing information on laboratory rats, mice and birds by the USDA decision. Because USDA does not collect information on the animals and report it to Congress, the group is unable to pass the information on to its members, and is prevented from carrying out one of its primary functions, the court decided.

Known as 'informational standing', the ruling appears to open the door to other lawsuits by animal-welfare organizations. Many previous legal challenges by activists failed when the organizations were unable to prove to the court's satisfaction that they had been materially damaged by government action, an essential requirement for legal standing.

Other animal organizations, including the Animal Welfare Institute, are closely watching the suit, with an eye on the newly released USDA regulations on primates, dogs and cats (see *Nature* 349, 641; 21 February 1991). Activists believe that the new rules, which give laboratories great flexibility in determining the details of enclosures and exercise regimes, are in violation of congressional intent. Now, they are considering suing.

The district court will turn its attention next to the merits of the animal defence fund's case. If it wins outright, USDA could be forced to include rats, mice and birds in its regulations. That would more than double the number of facilities the agency must inspect, and could cost several million dollars each year in additional regulatory expenses, according to USDA calculations.

Christopher Anderson



Is this laboratory mouse an animal? USDA says no.

Italy looks to next supernova

Mont Blanc

FOR Oscar Saavedra and his colleagues at the Italian national research council (CNR) Istituto di Cosmogeofisica in Turin and the Soviet Academy of Sciences, the 1987 supernova in the Large Magellanic Cloud came eight years too early.

Four years ago, the Italian/Soviet collaboration was the first to report a burst of neutrinos from SN1987A, which they had detected at their underground observatory at Mont Blanc, 2,000 metres below the Italian/French border in an emergency bay off the Mont Blanc road tunnel. That neutrino sighting, however, is now thought by most physicists to have probably been due to background noise and the detectors apparently missed a real neutrino burst that occurred several hours later.

So now INFN, the Italian national institute for nuclear physics, is building a larger neutrino detector at the Gran Sasso underground physics complex about 650 km southeast of Mont Blanc. This Large Volume Detector (LVD), which is expected to be completed in three to four years, might well have helped to establish whether the Mont Blanc detection was a real neutrino event linked to SN1987A if the LVD had existed at the time, Saavedra says.

Standard astrophysical theory holds that a collapsing star, before it explodes as a supernova, should release a burst of neutrinos, and the Mont Blanc team, headed by Carlo Castagnoli, believed that they had discovered such a burst. They were the first to report neutrino detection after other astronomers had reported optical observation of the supernova on 24 February 1987. But four and a half hours after the Mont Blanc burst, which consisted of five events over several seconds, a second burst was recorded independently by the Japanese detector at the Kamioka zinc mine and the US Irvine-Michigan-Brookhaven (IMB) device, located in an Ohio salt mine.

Physicists disagree on what the detection of two distinct signals implies. Because theory predicts only a single neutrino burst from a collapsing star, and because the IMB and Kamioka observations fitted well with theoretical predictions, the general perception among astrophysicists is that the Mont Blanc burst was background noise, most probably caused by penetrating radiation from the surrounding rock. Such a burst would be expected about once every three years simply from random fluctuations.

Saavedra believes it is possible that both the Mont Blanc detection and the later bursts recorded simultaneously in the United States and Japan could have been genuine events linked to SN1987A. This would require the star to have collapsed initially to a neutron star, releasing low-energy neutrinos picked up at Mont Blanc, but below the energy threshold of the IMB and Kamioka devices.

A second collapse to a black hole would then explain the neutrino burst recorded by IMB and Kamioka. At Mont Blanc, this burst may have been indistinguishable from the background noise.

Whatever happened, the combination of the Mont Blanc neutrino detector and the Gran Sasso LVD now under construction should give astrophysicists a clearer picture of the next nearby supernova.

Like the Mont Blanc device, the LVD will be a liquid-scintillation detector. The Mont Blanc detector contains 90 tonnes of an organic solvent held in 72 separate 1.5 m³ tanks. The solvent fluoresces when neutrinos, cosmic muons or other particles pass through it, and sensitive photomultipliers

OCEANOGRAPHY

Europe in the vanguard?

Turin

EUROPE should aim to take the lead in oceanographic research in the 1990s, says Umberto Colombo, president of the European Science Foundation (ESF), summing up last week's meeting in Turin on Oceans, Climate and Man, organized by the Foundation San Paolo.

The difficulties are several, including finding funds for oceanographic studies and turning out enough trained scientists, but the rewards would be worth the effort. Given the rapid expansion of oceanographic data and improvements in supercomputing expected over the next decade, many delegates at the Turin meeting argued that it may soon be possible to provide operational forecasts of the behaviour of the oceans, similar to the atmospheric forecasts used in medium-range weather prediction.

Ocean forecasts would provide information useful for shipping, fisheries, pollution control and coastal protection.

Added to this are the spin-offs for climate prediction: the difficulty of including the behaviour of the oceans in general circulation climate models is one of the most serious constraints on the reliable prediction of climate change.

The obvious model for a European ocean forecasting centre is the European Centre for Medium-Range Weather Forecasting (ECMWF), at Reading, England. Since 1979, ECMWF has been providing medium-range (three- to ten-day) weather forecasts for the meteorological agencies of its 18 subscriber countries, and has been a "roaring success", according to John Woods, director of marine and atmospheric sciences at the UK Natural Environment Research Council. But a parallel centre for ocean forecasting is still several years off, he says. The idea is one of many being considered by the European Committee on Ocean and Polar Science, a joint initiative of the European

around the tanks detect the emitted light, signalling the passage of the particle. The LVD is planned to be a scaled-up version of the Mont Blanc detector.

Some workers in the field say, however, that the Italians would have been better advised to abandon liquid-scintillation technology for their next detector. With a bank of relatively large scintillation tanks such as those planned for Gran Sasso, it is difficult to track the progress of particles through the detector. On the other hand, the devices at IMB and Kamioka, which work by detecting the Cerenkov radiation given off by particles passing through water, do allow researchers to follow the tracks of particles. This makes it easier to distinguish between different types of particles, and to tell if some events are simply due to faults in the equipment.

Peter Aldhous

Communities (EC) and the European Science Foundation (ESF), to present to ministers from European governments in 1994.

Nadia Pinardi, from the Italian national research council (CNR) Istituto per lo Studio delle Metodologie Geofisiche Ambientali at Modena, says that there is now enough data collected every 15 days to feed into an ocean forecasting model. The resulting forecasts would be coarser than those produced by meteorologists (ECMWF bases its forecasts on the daily collation of atmospheric data), but would be a useful start, she says.

Pinardi also wants the EC's research programmes to provide grants for European scientists to take part in the World Ocean Circulation Experiment (WOCE) — the huge international oceanography project planned to run until 1997 as part of the World Climate Research Programme — and for European researchers to take part in processing the large volume of oceanographic data expected from the European Space Agency's ERS-1 remote-sensing satellite, due for launch next month.

Although the EC's marine science and technology (MAST) programme has been valuable in stimulating European oceanography, Pinardi says, the EC has so far not paid for WOCE research. This is not so much a problem in countries such as Britain and Germany, where there are national programmes of WOCE research. But in most southern European countries, where national governments have shown little interest in WOCE, EC funding is the only hope for oceanographers who wish to participate in the largest research project in their field.

The biggest single obstacle for the progress of oceanography in Europe is under-training, says Woods. In Britain, for example, only ten per cent of scientists entering the field have previous specialist training in oceanography

Peter Aldhous

A foray into red tape

Tokyo

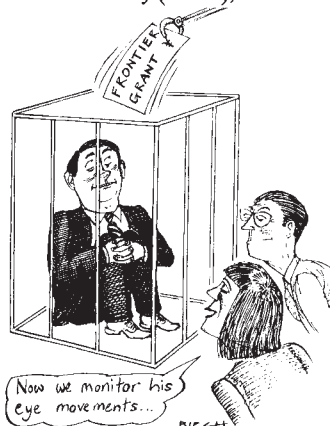
MORE than a hundred scientists from around the world have just won grants under the Human Frontier Science Program (see below), and for all but one of these the hard part is over and all that remains is to spend the money. The lone exception? A Japanese researcher who faces a still unknown number of hurdles before he can put his award to work.

Kenji Kawano of the neuroscience section of the Electrotechnical Laboratory (ETL) in Tsukuba, who will study the neural mechanisms of vision in collaboration with researchers in the Netherlands and the United States, has the dubious privilege of being the first person from a Japanese national laboratory to get a Frontier grant. As such, he will have to blaze a trail through bureaucratic red tape to get his award. Kawano's progress will be watched closely by hundreds of his colleagues in Japan's national laboratories who have hesitated to apply for Frontier grants because of the bureaucratic problems they face if they win an award (see *Nature* **349**, 185; 1991).

The problem confronting Kawano is that at present there is no mechanism for his Frontier grant to enter the accounting system for his institute. Japan's national laboratories, which include all of the country's government-funded research institutes except university-related institutes affiliated to the Ministry of Education, Science and Culture, are all fed from the same general pool of funds. If Kawano's grant is put into this pool, two things will happen. First, the grant will lose all identity and no longer belong to Kawano. And second, because of the rigid ceiling placed on the general account by the Ministry of Finance, funds would have to be displaced and eliminated from elsewhere in the account to make room.

That a Japanese researcher is subject to this Catch-22 is particularly ironic because it is Japan that started the Strasbourg-based Human Frontier Program and Japan still contributes the lion's share of the programme's funds — and these contributions come out of the same general account into which Kawano's award would be funnelled back.

Officials at the Ministry of International Trade and Industry (MITI), to which ETL is



affiliated, think they have a solution. They have told Kawano to accept the grant into his own personal bank account. This approach became practical only late last year after the Ministry of Finance introduced special regulations to exempt Frontier grants from personal income tax, as a way to deal with similar bureaucratic problems facing university researchers who won the coveted awards (see *Nature* **347**, 703; 1990).

Under a new set of special regulations just drafted by MITI, Kawano will then be able to buy equipment and supplies with his grant and bring them into his institute — after gaining the official approval of ETL's director by filling in a new set of forms. (Research-

ers were not previously allowed to bring such 'personal' items into the laboratory.) When Kawano no longer has a use for the equipment, he can 'donate' it to ETL by filling in yet another set of forms.

Kawano says he is reasonably happy with this proposal but is apprehensive about how it is all going to work. One concern is that he needs someone to feed the monkeys for his experiments on the neural mechanisms of eye movement. But under MITI's proposal he cannot employ a part-time assistant with his Frontier grant because an assistant does not qualify as 'equipment' or 'supplies'. However, Masayoshi Hamada of MITI's Agency of Industrial Science and Technology says that Kawano can get around this problem by employing a student under an existing system for inviting 'visiting scientists' to ETL. All Kawano has to do is claim that the monkey feeder is a 'capable researcher' who deserves to be invited into the institute.

MITI officials are clearly on Kawano's side and will do everything possible to smooth his journey through the unknown bureaucratic jungle that awaits him. And in the end, he may end up better off than researchers who receive equal-sized grants from other sources.

Because the money will be deposited in his own personal bank account under his own management, Kawano will not have to pay ETL any overhead from his grant. Normally, researchers at ETL and other national laboratories have about ten per cent automatically lopped off their grants to cover such things as gas and electricity bills and administrative costs.

Kawano should also be able to earn several thousand dollars in interest on his grant while it sits in the bank. MITI officials say it is entirely up to him what he does with the extra money. After all, with his extra duties as trail-blazer and guinea pig, he will have earned it.

David Swinbanks

US leads race for Human Frontier grants

Tokyo

JAPAN pays most of the bill, but it is US researchers who come off best in the allocation of grants for international research on the brain and molecular biology announced last week by the Human Frontier Science Program Organization (HFPSO) in Strasbourg.

US scientists lead 15 out of the 32 successful international research teams, which will get grants ranging in value from about \$500,000 to \$900,000 for three years. And 46 out of the 139 scientists in the teams are from the United States, far more than from any of the other participating countries.

The dominance of the United States shows the growing interest of US scientists in the Japanese-inspired programme. Eighty out of the 239 principal applicants for grants were from the

United States, more than twice the number from any other nation.

HFPSO was set up under Japan's initiative by the Western summit nations (Canada, France, Italy, Japan, the United Kingdom, the United States and Germany) and the European Communities in 1989. Switzerland also recently became a member. But Japan still provides more than 75% of the annual budget of about \$26 million for the programme.

The United States, on the other hand, has made no solid commitment to contribute money to the programme, although an official of the National Institutes of Health (NIH) recently wrote a letter to Sir James Gowans, the British secretary-general of HFPSO, promising 'in kind' support (see *Nature* **350**, 97; 1991); this support would apparently add up to a few hundred thousand dollars a

year, spent in the form of salaries paid to Frontier fellowship winners who go to institutes affiliated with NIH. US government officials argue that the United States already does a tremendous amount to support international research on the brain and molecular biology through NIH and other US-funded grant programmes and so does not need to contribute much to Frontiers.

Japan fared reasonably well compared with the other non-US member countries. Japanese scientists lead three of the successful teams and there are 27 Japanese among the grant recipients. But the number of Japanese principal grant applicants was greatly reduced this year — to 29, compared with 55 last year. The fall-off in applicants is apparently due in part to fear of bureaucratic barriers that face successful applicants in Japan. **D.S.**

Questions on drug pact

Boston

THE Harvard-affiliated Dana-Farber Cancer Institute has signed a cooperative research agreement with a leading pharmaceutical company that is raising anew questions about just how closely a university should collaborate with commercial interests.

In the arrangement, announced last month, Dana-Farber will receive up to \$100 million over the next ten years from Sandoz Pharmaceuticals Ltd, the US subsidiary of the Swiss company. The two will cooperate to develop anticancer drugs based on increased knowledge of cell signal transduction. Many researchers in the field believe that they now understand cellular signalling processes well enough to begin trying in earnest to develop drugs that will target specific signal transduction defects — problems in cells' chemical signals that give rise to the unregulated growth of cancer cells.

The agreement, due to begin on 1 January 1992, marks one of the largest investments so far by a pharmaceutical company in an academic research and teaching institution. But perhaps the most noteworthy feature of the plan is the way it is designed to pair academic researchers and corporate chemists in what David Livingston, vice president of Dana-Farber calls a "joint discovery" process.

The Dana-Farber Institute will devote roughly 10 per cent of the work time of its scientific staff to Sandoz projects. In addition, the arrangement calls for about 100 Sandoz chemists to work at the Dana-Farber side by side with its more than 400 research-

ers and postdoctoral fellows. Finally, in a departure for Harvard and other universities involved in large academic-corporate arrangements, patent rights and royalties from discoveries in the new programme will be shared by the two institutions rather than held by the academic institution and licensed to the company.

Livingston says he believes that the applied nature of the collaboration will fit easily within Dana-Farber's mission, because development of cancer treatment "is part of our mandate as a cancer center". But some critics worry that the arrangement could eventually weaken the institute's commitment to basic research. And Richard Kolodner, who conducts basic research at the Dana-Farber and chairs its division of cellular and molecular biology, says that many "legitimate questions" exist "about how something like this might influence hiring, or shift our mission."

As Kolodner observes, the Sandoz arrangement is unabashedly "goal-oriented", a characteristic that he said at least semantically separates it from previous collaborative plans involving Harvard. "Anyone here who wants some of this money will have to look at how their research might contribute to the development of drugs," Kolodner notes.

Such a feature could skew the direction of the institute's research, he says, but "it could also stimulate people to think about their research in new ways." All in all, Kolodner remains sanguine. "If you're at a cancer center, you can't help but think about the central problem at hand — treating and curing the disease."

Seth Shulman

PHARMACEUTICAL PATENTS

EC row looms over patents

London

A EUROPEAN Commission proposal to extend the life of patents on new drugs to a maximum of 16 years after a drug's commercial launch (see *Nature* 344, 482; 1990) has been rejected by the British government, to the annoyance of UK pharmaceutical companies. The British rejection of the plan, which aims to bring European drug patenting in line with US and Japanese practice, means that a row is likely when ministers from European Communities (EC) member states meet to discuss the proposal later this year — France and Italy have already passed their own laws based on the Commission's ideas.

The problem facing European pharmaceutical companies is that drugs must be patented before clinical trials begin. By the time a drug is licensed for commercial use, a 20-year patent often has only eight years left to run. After this time, cheaper 'generic' copies of the drug can be sold, made by companies that have not had to shoulder the burden of high research and development

costs.

The UK government says it will support the extension of drug patents, but only up to 13 years from a drug's launch. (The maximum effective length of patent protection in the United States is now 14 years, and in Japan 15 years.) The Commission's plan goes further than is necessary to help European companies compete with their US and Japanese competitors, said industry minister Lord Hesketh to the House of Lords, and the effect would be to discriminate against generic companies, and increase the costs of drugs.

However, the Association of the British Pharmaceutical Industry, which includes generic manufacturers among its membership, opposes the government's stand.

Apart from Britain, most of the other EC states with large pharmaceutical industries are in agreement with the Commission's plan, although the less-industrialized countries of southern Europe are expected to oppose the extension.

Peter Aldhous

Census sex puzzle

New Delhi

INDIA added another 160 million people in the past ten years to bring its population to 844 million. According to the 1991 census, the growth rate decreased by only 1.2 per cent compared to the previous decade.

Demographers are puzzled because the ratio of females to males in the population has declined, following an upward swing measured in the 1981 census. There are now 929 females for every 1,000 males, as compared with 934 in 1981.

At the turn of the century, India had 972 females for every 1,000 males, but the ratio had been steadily falling until the 1981 census. The reasons for the uneven sex ratio included a preference for male children, resulting in neglect of female babies, as well as a relative gap in health conditions between the two sexes — factors that many hoped had started to change.

The fact that the female/male ratio has fallen back since 1981 despite numerous women-centred health and nutrition programmes has set off a debate. Feminist groups believe that one cause is the selective abortion of females after sex selection tests offered by private clinics, which are mushrooming in cities and even small towns. Demographers, however, say the practice is not so widespread as to affect the sex ratio, and have called for an in-depth study to look for other reasons.

K. S. Jayaraman

SOUTH AFRICA

Saving the shark

Johannesburg

SOUTH Africa is to become the first nation to give official protection to the great white shark, *Carcharodon carcharias*. The Minister of Environment Affairs, Louis Pienaar, announced on 11 April that regulations to protect the species in South African waters will be promulgated shortly.

The legislation was necessary, Pienaar said, because international trophy hunters are starting to focus their attention on the South African coast. The shark's jaws command up to R12,000 as trophies, and populations of the predator have been decimated in some parts of the world.

In addition to making it illegal to catch great whites without permits, the legislation will forbid the sale of the species.

In future, permits will be issued only for research purposes or to the Natal Sharks Board, which is responsible for catching approximately 50 juveniles of this species each year in the shark nets that protect bathing beaches. The effect of the conservation measures on the local great white population will be monitored by Dr Len Compagno of the Shark Research Centre at the South African Museum, who is conducting research on its ecology.

Michael Cherry

Computerized grants project is unveiled

Bethesda, Maryland

THE halls of the National Institutes of Health (NIH) are paved with grant applications. So are the offices. Come deadline day, when 35,000 applications come in at the last minute, the Federal Express trucks are gridlocked in the parking lot unloading funding appeals. By the time NIH finish making 75 copies of each application and distributing them to all the appropriate offices, the agency is buried under more than 2.5 million applications, each about a hundred pages long — over a thousand tonnes of paper.

In the electronic age, this flood of paper is an embarrassing thorn in the side of the NIH and other US science agencies. Years of vague proposals to computerize the process have produced some abortive trials, several committees and plenty of excuses.

But now, a handful of computer evangelists with some hot machines are finally making progress towards the grail of grant-making: the all-electronic application and review. No paper, no post — nothing but bits and bytes.

Earlier this month, grants officials unveiled their plans for an electronic future at a meeting of the Division of Research Grants advisory board attended by senior NIH staff. Experimental grant software is nearing completion, they said, and will be tested in preliminary trials in early fall at the University of

South Carolina, the University of Washington and another, yet unnamed, university.

In addition, a collaboration with the National Science Foundation (NSF) is aiming at a common applications form and software that could be used at both science agencies. The two combined receive some 70 per cent of all grant applications to the federal government.

Researchers are making progress towards the grail of grant-making: the all-electronic application and review. No paper, no post — nothing but bits and bytes.

John Mathis is the motive force behind the most ambitious of the NIH projects — one that will eventually not only allow scientists to create and submit their proposals electronically, but will also let NIH computers read and categorize them without human help. Mathis's vision of a computerized grant-making machine is a mix of simple personal-computer software and complex artificial-intelligence (AI) programs.

Known as the EGAD (Electronic Application and Grant) project, the effort is about to show its first face — a software program for distribution to universities that will check applications for completeness and accuracy in real time as a researcher enters the infor-

mation.

The software (designed for Macintosh or IBM-compatible computers) will make sure that the numbers add up at the bottom of the financial portion, and that the researcher has not accidentally given something like 'biology department' as his or her last name.

Typing time and labour are just the most obvious of the savings. Mathis hopes that a sophisticated AI program that is still under design will also take over the tedious chore of assigning applications to the appropriate NIH institute and study section, a task that now requires a panel of in-house scientists.

The AI program, known as 'the referral assistant', will scan an application, looking for information that reveals its scientific field and subject. Several techniques for doing this are being considered, including keyword searches and citation analysis, which would use the references cited at the end of the application as a guide to the science within.

At the National Science Foundation, however, progress on an electronic grants package has been tempered by a sobering example of how hard the transition can be.

In 1988, NSF started a project known as EXPRES to produce a program that would allow scientists to create and submit grant data electronically. Two universities, Carnegie Mellon and the University of Michigan, were to develop a standard 'document

Electronic review: mixed messages

Washington

SCIENCE agencies may save a lot of paper-shuffling by receiving grant applications electronically, but the real delay and expense is still the review process after the applications arrive. Both the National Institutes of Health (NIH) and the National Science Foundation (NSF) fly hundreds of researchers to Washington each year to staff panels that review stacks of applications, ranking them in order of scientific interest. The entire review process costs the two agencies some \$80 million each year, and can take up to nine months to complete.

Both agencies are considering replacing the traditional face-to-face review group with their electronic equivalents. But as a few trial projects have shown, getting scientists to give their full attention to peer review while sitting in their own offices is a challenge that may defeat even computers.

In an experiment conducted last year, NSF submitted 52 applications to electronic peer-review. After the agency mailed the applications by regular post, the panellists submitted their reviews and exchanged comments over Internet, a nationwide computer network.

Overall, the experiment was generally a success. NSF found the comments of

the 'e-panels' to be more thoughtful and the reviewers more prepared than in similar face-to-face panels. But they also noted that interaction among the panel members ranged from minimal to virtually non-existent.

"There was very little exchange of e-mail," one officer wrote in a NSF report on the project. "In one panel it was because the opinions were unanimous. The other panel was a disaster. I couldn't get them to do much of anything."

Reviewers, of course, liked not having to travel to Washington, DC. (So did NSF. Bringing a scientist to headquarters costs the agency about \$1,000 per day. NIH spend more than \$20,000 for each of their three-day study sections.) Most reviewers also said they could live without the body language and verbal clues of a face-to-face panel. For small (3-4 people) groups with three to six proposals each, the extra time to prepare thoughtful comments and opinions was worth the lost personal interactions, they said.

NSF, which tends to have smaller review groups and to give programme officers more say over funding decisions, may eventually be able to adopt e-mail reviews for the two-thirds of all reviews that are now conducted at least partly

with face-to-face groups.

The NIH, on the other hand, depend almost exclusively on their study sections, which usually number about 20 people — probably too large for e-mail review. For some special reviews, however, NIH assemble panels of as few as five reviewers. In past experiments, NIH have occasionally convened panels by conference call and — in at least one abortive trial — mail.

The agency found that conference calls appear to work relatively well at a pinch, although they have none of the advantages of face-to-face reviews (the nonverbal communications) or e-mail reviews (the more careful consideration). But getting reviewers to respond by mail turned out to be hopeless. Almost all the comments came in late, and some never came in at all. "We had to throw some reviews out altogether because we didn't get even three responses," says John Mathis, who ran the NIH experiment.

Until researchers learn to type as fast as they can talk, they are likely to resist being handed a stack of grants for electronic review. Although the trend towards increased use of e-mail is clear, for now, most reviewers would rather fly to Washington than write a ream of comments about somebody else's grant. **C.A.**

TELECOMMUNICATIONS

Satellite aids sore fingers

Munich

architecture' or text format, as well as two different word processor programs that could understand the format. But the universities never reached the point at which their software could share files, and the project was killed last year.

The only concrete product to come from the NSF project was a simple text processor called PS-Express, which is now widely distributed among scientists. It converts a simple list of data elements (such as 'Last Name = Mathis') into a file written in the common Postscript graphics language. When sent to a Postscript-compatible printer, the file prints out an entire NSF application form with all data filled in.

NSF also accepts PS-Express files electronically, over the Internet computer network. Some 200 applications have been filed that way in the past two years.

But electronic submission at NSF "is still a very hand-held process," says Lawrence Edwards, NSF senior project manager for office information systems. "When we get an application electronically, I print it out, take it downstairs and put it in the queue, just like it came in the mail."

NSF is now developing a simple 'extract' program that will read the PS-Express files and automatically transfer 20–30 essential data elements into the main NSF database, without the need to type it in. That saves time, and more importantly, errors.

But more ambitious projects are in abeyance until the NIH experiments prove that researchers, administrators, agency officials and the technology are all ready for the leap.

So far, government acceptance of the electronic grant projects has been cautious — for good reason. Although the technology has come far in the past ten years, most computer monitors are still barely adequate to display an entire application page at once, clearly or not. "I'd hate to read a proposal on the screen," Edwards says.

And at NIH, Mathis is still fighting computer phobia. "We're waiting for a generational change," he says. "People keep telling me, 'John, what you're doing is terrific. I just hope I retire first.'"

Mathis takes the long view. "Everybody knows that (electronic grants) are inevitable," he says, "it's just a question of when to buy in. If you want to be on the cutting edge and to have a lot of fun, you do it now. Otherwise, you step back and let everybody work the bugs out."

Bugs or not, cultural resistance remains the largest hurdle to electronic grants. Mathis and his staff spend much of their time visiting universities, other parts of NIH and federal agencies to preach the gospel of computerized applications. These "dog and pony shows", as Mathis describes them, are intended to create a groundswell of demand. Until Mathis and Edwards can convince university administrators that electronic grants save time and trouble, the weak link in the chain will continue to be the human one.

Christopher Anderson

As anyone who has tried in recent months to place a call from western Germany to eastern Germany can attest, the two telephone systems are anything but unified. A re-dial button or a rubber finger is a virtual necessity, since the few available lines are always busy.

But now the national telephone company Telekom is bringing in high technology to widen the bottleneck in at least a few geographic areas until a more thorough upgrading of eastern Germany's antiquated system can be carried out. Earlier this month, Telekom began switching calls between some eastern and western cities by satellite in an efficient, if expensive, attempt to alleviate the situation.

From 5 April, Telekom made available 30 additional satellite lines between the western city of Hamburg and the eastern city of Dresden to complement the paltry 24 land lines previously available.

Hamburg has a population of more than 1.6 million; Dresden has 500,000 residents and is the third-largest city in eastern Germany. An additional 36 lines are expected to be added later this month.

It is the first time a satellite has been used to route public calls between the two parts of Germany. The satellite, known as Kopernikus, is normally used for intercontinental calls.

But the stopgap use of the satellite does little to ease the misery of private and corporate customers in other regions of Germany. Telekom, and the Ministry of Post and Telecommunications to which it belongs, have been lambasted in the German press for their sluggish response to the immediate need for more telephone lines between the two regions of Germany. As Telekom has a monopoly on offering basic telephone service, there is nowhere else the customers can turn.

Telekom is planning to invest DM55,000 million (about \$33,000) by 1997 to rebuild the neglected eastern German telephone infrastructure. A temporary network, known as an overlay network, is expected to be in place by July. It will be used as an adjunct to the existing network in eastern Germany, most of which was installed back in the 1920s.

Steven Dickman

MONEY MATTERS

Gaussian curve graces banknote

Göttingen

THE German mathematician and astronomer Carl Friedrich Gauss (1777–1855) is honoured on a new German 10-mark banknote issued on 16 April in Göttingen. The bill marks the second time in as many years that a German banknote has been issued featuring a scientific luminary. Last year, immunologist Paul Ehrlich was so honoured (see *Nature* 347, 415; 1990).

Gauss, for whom approximately 50 mathematical laws, formulae and methods have been named, was head of the Göttingen astronomical observatory from 1807 to 1855. One of the founders of modern geophysics, Gauss became famous during his lifetime for his method of measuring magnetic fields in absolute units, which later became known as gaussian units, including the gauss (for magnetic field strength).

In addition to a portrait of Gauss, the face of the note features the familiar gaussian distribution. The reverse side features a 'viceheliotrope', a type of sextant used for surveying invented by Gauss to divert a ray of sunlight to a distant observer. Circles in the background represent stylized plane-



Gauss is latest in the line of scientists to receive recognition on German currency.

tary orbits and magnetic fields.

Gauss is latest in the line of scientists to receive recognition on German currency.

In 1992, a third figure from German scientific history, this time a woman, will appear on a new 500-mark note: Maria Sibylla Merian (1647–1717), who compiled a famous catalogue of drawings of insects. The series of notes is designed to depict great personalities of German cultural life.

Steven Dickman

Two views of the new Germany

SIR — The comments of Frits L. Meijler (*Nature* 350, 268; 1991) on the purported German character are hurtful in two ways: (1) It is horrible to see how deeply he has been hurt, and still feels; (2) I resent his presenting a harsh verdict, then acknowledging that E. Guthy (*Nature* 348, 670; 1990) did his best but claiming that Guthy did not have a right to say what he did because he is German. In effect, the accused is not allowed to defend himself because he is being accused.

Meijler is a generation older than I am, judging from his letter. For me, things do look different. It is my firm belief that, if there is a common strain among Germans born after 1950, it is a constant if unexpressed awareness of Auschwitz. It has pervaded the political thinking of my generation to such an extent that many foreigners consider its effects rather strange. A friend of mine once remarked, "What Germans cannot understand is that I can say, I am an American, and I just don't feel bad about it". I learned only when I went to the United States that other peoples do not have a taboo on national feelings. Germans do, even after 9 November 1989, because 9 November 1938 is by no means forgotten. If there are some groups of thoughtless people, such as those who threw stones at Poles when the border was opened for them this week, they are a reason for concern, like the soccer hooligans in Great Britain; but if there is anything typical in the incident, it is the immediate sense of alarm in the public reaction.

If Meijler related his personal feelings on that matter, may I relate mine? How it is to grow up, becoming slowly aware of a horrible past, learning that there is a cause of great shame although I had not done anything wrong? Knowing that my father fought in Hitler's army, from loyalty to his country, as is today considered normal in any other country but Germany? I left Europe, in part to leave all this behind. To no avail.

Around the age of 15, I was told that one can tell a Jew by the look on his face. I was outraged, thought this to be a racist opinion. Then in California I met a very beautiful young woman with a face unlike any I had ever seen before. One day I saw in a bookstore a book of photographs, apparently taken by an SS officer at the ramps in Auschwitz. There were long columns of people, bearded men with hats, then a close-up of the first row of the women — and there was my friend's face, six times, with the knowledge of death in their eyes.

Nothing reveals the backward direction of Meijler's view more than the use of the phrase "Grossdeutschland". All right, Germany won the soccer championship last year, so what? No peoples are more eager to get along with their neighbours than Germany, precisely because of what happened to Meijler's parents. Regaining a sense of national feeling is probably a healthy process.

When the borders opened on that day in November, I felt that an incredible burden had been taken from me which I had been carrying for so long that I did not feel it any more. I was utterly surprised at myself. What I observe now among the public is not an exaggerated sense of pride, but rather a sense of being 'all right' again.

A more genuine reason for concern now perhaps is that the former East Germans are not used to dealing with people from other nations. They definitely react in a different way from Westerners to Turks, Yugoslavs and Poles. They will have to learn.

FALK KOENEMANN

*Institut für Mineralogie, RWTH,
D-5100 Aachen, Germany*

SIR — I read with interest the striking article on "Germany in perspective" by E. Guthy about a revival of prejudice against Germans based now on German "national or racial characteristics". It is surprising but true that at this stage of development in genetics such articles are not only desirable but indeed very much needed. Even among scientists there are still some who tend to forget that neither a German nor for that matter a Jewish gene was ever sequenced.

During the pre-Nazi and Nazi period, the most famous champions of the Nazi 'genetics' were the German university professors E. Fischer and Othmar von Verschuer. They accepted as a dogma the superior 'national and racial' characteristics of Germans and taught that the social, cultural and political activities of some religious and/or ethnic minorities are genetically predetermined and therefore forced sterilization or 'elimination' (that is, robbery and murder) of certain minorities is justified. Now, apparently, some try to explain German politics in the same but reverse vein by assuming the existence of a malicious 'German gene'. It would appear that the Fischer-Verschuer obscurantism in genetics (and that of their ambitious postdoc Josef Mengele) is still alive and kicking even among those with a scientific background. After all, the very same von Verschuer remained for almost 20 years an active university professor of 'anthropology' and genetics in post-war Germany (cf. Benno Müller-Hill *The Murderous Science*, Oxford University Press, 1989). For such reasons Guthy's wise words should be doubly appreciated.

On the other hand, I share the concerns of W. Frank Harris and of J. F. MacKenzie (*Nature* 347, 510; 1990) that, with the new economic power and without the constraints of partition, the destructive and inhuman forces of the past might loom up in the post-Bittburg Germany, although I would not look for their origins in a 'German DNA'. It is true, for instance, that Chancellor Helmut Kohl organized a wreath-laying ceremony on SS graves in Bittburg in the 1980s, but this was not genetically predetermined.

Many other frightening academic and political phenomena (as the Verschuer and Bittburg cases) were tolerated during the four-and-a-half decades of post-war developments in Germany. One can therefore understand the concerns (mentioned in your leading article in *Nature* 346, 203; 1990) of Mrs Margaret Thatcher and the group of academic historians, who "wondered whether some of the unhappy (German) characteristics of the past might reemerge in 10 or 20 years with just as destructive consequences". It is not irrelevant in this connection that the facilities for manufacture of poisonous gas in Libya, Iraq and Syria were provided by (among others) contemporary German technology and business, albeit run by indigenous talent. I do not accept the theory that this could have been done while the German authorities were sleeping. Neither do I believe that a 'German DNA' is responsible.

The differences between Guthy's and Mrs Thatcher's perspectives on Germany arise from a different approach to the problem. Guthy in his article covers centuries, whereas Mrs Thatcher is influenced by living memory. I gladly agree with Guthy that the past centuries of glorious German history are a guarantee for the future glorious centuries. The question remains of what is going to happen in the next 10 to 20 years. There are many good reasons for optimism as well as for anxiety and vigilance.

RUDOLF VRBA

*University of British Columbia,
Department of Pharmacology &
Therapeutics,
Faculty of Medicine,
2176 Health Sciences Mall,
Vancouver, BC, Canada V6T 1W5*

Mix-up

SIR — I read with interest the book review by Joseph F. Sambrook about genomic genomes (*Nature* 349, 752; 1991). My father, Lionel Penrose (1898–1972), geneticist and friend of Frances Partridge, might well have felt flattered to be taken for Sir Francis Galton (1822–1911), who was a founder of human genetics.

It was only in 1956 that the number of human chromosomes was clearly shown to be 23 pairs, and the chromosomal abnormality in Down's syndrome was found to be trisomy 21 in 1959.

SHIRLEY HODGSON

*Guy's Hospital,
Division of Medical and Molecular
Genetics,
7th & 8th Floors, Guy's Tower,
St Thomas Street, London SE1 9RT, UK*

Correction

In the letter entitled "Early case of AIDS in the USA" (*Nature* 347, 509; 1990), the authors should have been listed as both R. F. Garry (Tulane University) and M. H. Witte (University of Arizona). □

Does climate still matter?

Jesse H. Ausubel

We may be discovering climate as it becomes less important to well-being. A range of technologies appears to have lessened the vulnerability of human societies to climate variation.

AMIDST widespread agreement that the planet is committed to at least some climatic change induced by human activity, there is growing pressure to "slow the greenhouse express" (ref. 1). Here I examine climate through the lens of technology and innovation, to clarify what adaptations have succeeded and the trends in vulnerability to climate. I also examine whether the greenhouse effect by itself will call into play new technologies, or whether the evolution of technology will largely be 'business as usual' regardless of climate change. Finally, I identify some ways that government may assist adaptation.

My focus is on adaptability of human systems, including agriculture. Adaptability of ecosystems and the ethics of human behaviour that brings about large-scale transformations of the Earth must also be considered in balancing responses to the greenhouse issue.

What are often labelled adaptive measures are themselves the impacts of climatic change. Innumerable adaptations in food, clothing and shelter are responses to the spatial and temporal variability of climate. Humans do not wait guilelessly to receive an impact, bear the loss, then respond with an adaptation. Rather they attempt to anticipate and forestall problems.

Innovations

Technical innovations relevant to climate and their diffusion occur in all societies and sectors and in many forms. Technology includes both hardware (for example, materials, physical structures, devices and machines) and software (rules and recipes for behaviour).

Illustrative innovations of hardware are cisterns and dams to store water, tractors to speed rapid harvests, and new crop cultivars to reduce susceptibility to drought. Perhaps less obvious, but of great importance for adaptation to climate, are information technologies. In the United States, during several years in the 1980s sales of information technologies to the agricultural sector were comparable to sales of farm equipment. The long history of software innovations includes tide tables, irrigation scheduling, and weather forecasts. Along with the readily classified hardware and software are climate-related behavioural, social and institutional innovations, such as agricultural credit banks, national parks, green political parties and flood insurance.

Software and social innovations are almost always indispensable for the technol-

ogy of new hardware. Major innovations, in transportation for example, are in fact clusters of innovations involving not only new materials and physical processes but also new forms of social organization, including financing and management. Transportation systems exemplify technology that has been important in adaptation to climate through expanding the availability of food from a local to a global scale. Reliance on food from afar not only diversifies diet, but also spreads production risks across more climatic zones.

Early communities drew the bulk of their food from small areas. One of the earliest city states, Uruk in Mesopotamia, probably grew most of its sustenance except for animals within 20 kilometres of the city walls². Two millennia later, Greece and Rome obtained most of their food from overseas colonies across the Aegean and Mediterranean seas³. At its height, Rome acquired 200,000 tons of grain annually for its one million inhabitants, most of it shipped from Africa, Sardinia and Sicily. What marine transportation did in the classical world, the steam locomotive did in the nineteenth century, halving the cost of land transport. The railroads penetrated the great land masses of North America and Australasia. Their operations were little affected by climate, topography or other local conditions. Great parts of the new worlds were dedicated to cultivation of single crops to supply world markets and to smooth availability through the year.

Technological inventions and innovations that have had roles in human ability to adapt to climate over the last 100 years or so range widely⁴: food preservatives (1873) to overcome problems of seasonal food production; light bulbs (1879) to make work safe and effective indoors; aluminum (1887) and other structural materials to resist environmental deterioration; refrigeration (1895) and air-conditioning (1902, 1906) to facilitate activity in hot seasons and locations; automobiles (1890s) to provide personal transportation that is much less sensitive to climate than horses or pedal bicycles; mechanical windshield wipers (1916) to see in the rain; antifreeze (ethylene glycol) (1929) to safeguard motors in winter; frozen food (first sales, 1930) to diversify diet among regions and seasons; radio-beam navigation (1934) to fly in poor visibility; and weather (1960) and Earth resources (1972) satellites for analysis and forecasts of weather and climate.

In many respects we seem to be 'climate-proofing' society, making ourselves less subject to natural phenomena. For centuries and millennia we relied mainly on behavioural

and social adaptation. We took siestas when the sun was high and sought refuge in hill stations in the monsoon season. Large pastoral and nomadic populations followed the seasonal availability of resources and avoided climatic stresses. Much of the planet remained seasonally or entirely uninhabitable for climatic reasons. With current technology many people can live in virtually any climate that now exists. Modern water supply and heating and ventilation systems, along with medicines (for example, quinine and vaccines) and public health measures, have enabled large populations to inhabit formerly uninhabitable regions. By 1980 the population in semiarid, desert and mountain regions had passed 35 million or 15 per cent of the US population⁵. Lacking modern technology, these zones accommodated less than one per cent of the population in 1860 and six per cent in 1920.

Preferences

The ability to colonize almost the entire planet is due to the human ability to carry with us that particular range of environment in which we can survive and prosper⁶. In wealthier societies, preferences are shifting toward hot and dry climates that were forbidding a century ago. Evidence of lessening human vulnerability is also found in health. For example, a flattening of monthly rates of total mortality in Japan between 1899 and 1973 is explained partly by diminution of the previous, climatically driven seasonal peaks⁷.

Production can now proceed more or less continuously in severe environments. For example, North Sea offshore oil platforms operate 24 hours per day, 365 days per year. At a price, services, such as aviation, are now available at almost all times and places. Aviation began as a system that was extremely sensitive to weather and is increasingly less so, due to expanded range; avionics, radar and guidance systems; understanding of thunderstorms, wind shear, and other weather phenomena; and changes in construction materials. Crops and livestock can now be produced 'indoors' protected from the elements. In some cases, alternatives to outdoor production are so advantageous that a crop is displaced. Originally spurred by the need for supplies in wartime, synthetic rubber from petrochemical feedstocks, which is not subject to the vagaries of pests, droughts and floods, and other risks outdoors, has overwhelmed natural rubber from trees. In 1990, ten per cent of US fish production was in the controlled environment of

fish farms⁸, up from one per cent in 1980 and projected to reach 20 per cent in 2000.

Consumption is also insulating itself from environment. Inside most shopping malls, for example, only fashions or decorations signal the season. Sports are increasingly played in domed stadiums isolated from the weather. In affluent societies, winter vacations in warm climates have become a popular adaptation to escape climatic impacts.

Forecasting is itself a key technology that reduces vulnerability to weather and hence climate. Forecasts can help accommodate peak loads in electric power systems during heat waves. Improved forecasting, in conjunction with increased incomes and better transportation, has also enabled more people to enjoy recreation in all seasons.

Synchronicity

The decline of 'synchronicity', the naturally enforced time regimen of social groups, is a feature of life in advanced societies. In agricultural societies, the rhythm of life is strongly determined and coordinated for almost all by the seasons and the associated demands for labour in the field. In advanced economies, where both production and consumption may proceed almost continuously and only about five per cent of the population farms, weather and climate no longer control schedules. The fact that the peak season for holidays in advanced societies is the late summer, a peak season for labour in agricultural societies, indicates the transformation that has taken place.

In the late 1970s and early 1980s a group of US researchers explored the 'lessening' hypothesis of climate impacts⁹, which states that persistent and adaptive societies, through their technological and social organization, lessen the impacts of recurrent climatic fluctuations of similar magnitude upon the directly susceptible population and indirectly lessen the impacts on the entire society. In the cases studied, substantial evidence was found to support the hypothesis of lessening impacts. For example, in the US Great Plains, the most severe disruptions to livelihood and health occurred during the earliest periods, when incidences of malnutrition and starvation were recorded. Investigation of the more recent periods showed much smaller impacts for comparable drought stress, because of a variety of adjustments and strategies, including more extensive and effective anticipatory action.

An alternative to the lessening hypothesis is that increasingly elaborate technical and social systems insulate us from the adverse effects of recurrent climatic fluctuation at the cost of increased vulnerability to catastrophe from less frequent natural and social perturbations. In a global economy, such vulnerability might be devolved or shared ever more widely. Presumably this vulnerability to catastrophe, surprise or nonlinear effects is what worries many about the greenhouse effect. But the evidence seems to weigh against the suggestion that technology, lifestyles and

other forces are creating a more 'brittle' system in the face of climate change.

Society appears to proceed along 'technological trajectories' that enable, for example, more travel, more financial transactions, and more messages. The succession of technologies that make possible this increased activity appears to diminish in sensitivity to climate. Although it was usually true that neither rain nor sleet stayed the American postman from his rounds, no system of postmen could be as faithful as the modern telecommunications system that now carries a much larger share of messages than the old system of letters. Similarly, a system of energy from wood and hay was more climatically sensitive than one reliant on oil and natural gas. Water and wind power are, of course, more sensitive to climate. In the late eleventh century, the areas under Norman rule in England had about one water mill for every 50 households¹⁰, providing power to grind grains, work metal, weave textiles and cut wood. In 1694 France had 80,000 flour mills, 15,000 industrial mills, and 500 iron and metallurgical works, altogether almost 100,000 facilities powered by wind and water. Although such an industrial infrastructure is tightly adapted using climatic resources, it is also vulnerable to climate variability and change.

The trend toward less climatic vulnerability also exists in transportation. Well into the nineteenth century, sailing vessels were the preferred long-distance transport and frequently becalmed. World steamship tonnage exceeded sailing tonnage only in 1893. Although coal cost more than wind, steamships rapidly became cheaper as well as faster than sailing ships, because their schedules were more regular and avoided the circuitous routes required by sailing vessels. Transport underground through tunnels by high-speed magnetically levitated trains is already on the drawing boards in Japan; such systems would be less sensitive to climate than the surface and air systems now in use.

Climate is only one of several factors that have driven the evolution of systems of communications, transport and energy. It is probably secondary. But vulnerability to climate and other environmental forces may be a good proxy for quality and reliability of service. To the extent that the systems evolve in the direction of higher quality and reliability, these trajectories of development may also decrease vulnerability of major systems to climatic change. It would be useful to identify exceptions to this pattern, should they exist.

As incomes depend less on activities out of doors, societies become less vulnerable to climate. The trend in all developed countries since the industrial revolution began is away from employment outdoors (Fig. 1) and toward employment in the service sector, most of which is in climate-controlled office buildings and shops. With a lag of 50–100 years the same trend is found in less developed countries, where 40–50 per cent of the popu-

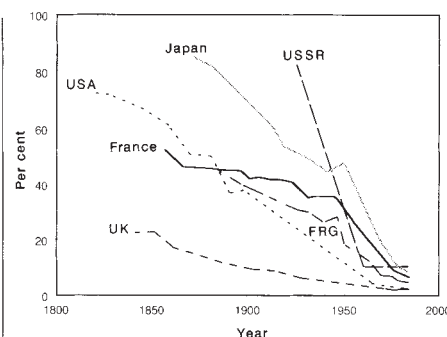


FIG. 1 Share of the workforce employed in agriculture²¹.

lation is now engaged in agriculture. If the current trend continues, this fraction should diminish to 20–25 per cent by 2050.

What are the times characteristic of technical innovation and diffusion of technologies in relation to the time of human-induced climatic change? Taking account of both CO₂ and non-CO₂ greenhouse gases, major climatic shifts are expected during 40–50 years (ref. 11). A retrospect on technology during the past century suggests the extent of change during the decades to come: in 1890 there was little farming in California and Australia, and key technologies did not yet exist or were not widely applied to impound and transport water, or to transport and store agricultural products promptly; in 1903 the Wright Brothers flew 59 seconds at Kitty Hawk, North Carolina, whereas in 1990 in the United States, over 500 million air passengers flew around 450 × 10⁹ miles¹²; commercial transatlantic aviation, relying on the jet engine, superseded travel by ocean liners about 1960; tourism has been extended to Antarctica, and scientific bases there are occupied year round, despite the fact that it was only in 1911–1912 that Roald Amundsen reached the South Pole and Robert Scott died there; penicillin, the first important antibiotic, was discovered by Alexander Fleming in 1928, and large-scale production began only as recently as 1943; nuclear power for electricity generation first came into use in the late 1950s, but within about 30 years it was able to provide over 70 per cent of France's electric power; until 1965 no satellites were used for any routine application, whereas satellite systems now girdle the Earth, watching for storms, relaying communications, and helping ships and planes navigate anywhere on the globe; and finally, although the microprocessor was only introduced in 1971 and the personal computer appeared about 1977, Americans are now using over 50 million personal computers.

During the period in question and with or without climatic change, technology will transform the way people live. Food, energy, transportation and all the other systems that support human life and the economy will be changed by technologies that can be glimpsed now, such as genetic engineering, fusion, superconductivity and desalination, as well as by technologies yet to be easily pic-

tured. Fifty to a hundred years will allow the replacement of most major technological systems. Indeed, 50 years are enough time to turnover almost the entire capital stock of the society. About two-thirds of capital stock is usually in machines and equipment and about one-third in buildings and other structures. In Japan, the average renewal period for capital stock in business, the time it takes for machinery and equipment in an industry to be almost completely replaced, ranges from about 22 years in the textile industry down to ten years or less in such fast-moving industries as telecommunications and electrical machinery (see Table). In agriculture the estimated life span of cultivars in the United States is seven years for maize, eight for sorghum and cotton, and nine for wheat and soybeans¹³, and most experts believe the life span of cultivars will grow shorter. Relative to greenhouse warming, turnover is also fast for nonmachinery capital stock, which includes buildings, pipelines, and so forth. According to recent surveys, in the Federal Republic of Germany in 1985 some 60 per cent of the stock is less than 20 years old and in the Soviet Union some 80 per cent is less than 20 years old (Fig. 2).

At first such figures may surprise us. But some reflection about the built environment relieves the surprise. Consider the office space in a modern city. Most of the space is in buildings completed in the last 20–30 years. These new buildings are filled with new equipment, for example, new telephone systems. Indeed, even older buildings are filled with modern equipment that did not exist 15 or 20 years ago. The same is true for supermarkets, restaurants and other stores. A large fraction of residential housing is similarly young and, in turn, filled largely with new domestic appliances of all kinds. If societies grow at a rate of 2–3 per cent per year, as the industrialized societies have for the past 150 years, then half of all capital stock will always be younger than 30 years old.

Probably the systems that take longest to build are infrastructures. Even these are constructed in less than a century¹⁴. Many infrastructure systems are (or should be) continuously reconstructed. For example, roads are repaved every 5–15 years, depending on use. The 7,000-kilometre canal system of the United States was almost entirely built in the 30 years between 1820 and 1850. More than 90 per cent of the 300,000-mile US railway network was laid in the 65 years between 1855 and 1920. The paving of virtually the entire six million-kilometre surface road system of the United States was accomplished between 1920 and 1985. The US interstate highway system was completed in about 30 years from the time that it was announced by President Eisenhower. It is interesting to consider whether climatic change could require any public works on this scale; coastal protection and interbasin water transfer would seem the most likely candidates.

Because capital stock continuously turns over on a time scale of a few decades, it will

CAPITAL STOCK RENEWAL	
Industry	Renewal period
All industries	13.4
Manufacturing (all)	15.8
Electrical machinery	9.8
Transport machinery	13.2
Pulp and paper	13.7
General machinery	14.2
Chemicals	16.6
Food stuff	16.7
Steel and non-ferrous metals	21.1
Textiles	22.5
Non-manufacturing (all)	11.8
Service	8.1
Transport and telecommunications	10.7
Construction	11.9
Finance and insurance	12.8
Distribution	15.6
Real estate	15.8
Electricity, gas and water supply	15.8

Average renewal period for capital stock of business corporations in Japan by industrial sector, 1986–87 (ref. 22).

be possible to put in place much technology that is adjusted to a changing climate. This can be done without extraordinary measures given reasonably accurate information about the future. For the shortest lifetimes, even accurate information about the present climate will do.

That perhaps 90 per cent of the global capital stock in the year 2040 will be built after 1990 does not diminish the significance of some long-lived structures. Action may be necessary to protect cities such as Venice, where preservation of historic buildings is the goal. In such cases, the process of replacement is not relevant.

The adaptations for long-term climatic change will probably mostly be the same as for other climate variation. The technologies, small and large, that buffer human activity over the long-term will be the same ones that mollify the difference between daytime and nighttime temperatures, protect against normal variability between days, shield from storms and hail, adjust to the seasons, and adapt to the wide range of climates where people already live.

No one has yet presented a radical innovation uniquely adaptive for the greenhouse effect. The main innovations directed at the greenhouse effect are so far organizational, in particular research groups in universities and government and assessment groups, such as the Intergovernmental Panel on Climate Change (IPCC). If the future greenhouse climate in any place will consist of climates that already exist somewhere on Earth, then many of the adaptations may look familiar.

Because the population of the world is imploding into cities, it seems logical that technologies that make cities habitable in unwelcoming climates will be among the technologies that are most important. Houston,

Phoenix and Toronto may offer lessons. The trend in such places is toward ever larger enclosures of space and passageways connecting them, where workers and shoppers are not subject to the elements. Already during much of the year in such cities few people are seen outside on the streets. Technologies for 'smarter' buildings and for more efficient building materials should be adaptive. Cities in developing countries, which are often in difficult climates to begin with and face worsening problems of urban pollution, may well lack the resources to apply such technologies to raise or maintain the quality of life in the face of changing climate.

Until this century much of the human struggle with climate was to keep warm. Because the struggle succeeded, in 1850 the population in Europe, a land of well-chronicled and damaging winters, was three times as large as that of Africa and nine times as large as that of Latin America¹⁵. Now a main change in adaptation will be emphasis on technologies to stay cool. There is already pressure and success in this direction, population grows in tropical regions and people migrate south in temperate zones. Chemicals for refrigeration that do not exacerbate the greenhouse effect may thus earn a premium, and, of course, low greenhouse gas emission technologies to produce electricity and energy in general.

For water resources, larger-scale control of flows may be the trend. The Thames Barrage and the Netherlands Rhine Delta scheme may exemplify massive hydraulic systems that will be imitated in areas of major coastal populations. Fresh-water systems in some regions would also be made more robust by extending networks of supply over wider spaces through interbasin transfer and other strategies. In practical terms, many of the technologies needed may be well-established, for example, tunnelling, pumps and other technologies traditional to civil and mechanical engineering, updated with electronic sensors and other devices for management and control. Technologies for management of water demand will be equally or more attractive in many regions; these would include not only hardware technologies for minimizing leaks, but also software technologies from operations research, as well as economic and other incentives.

In agriculture, with a few notable exceptions, most emerging technologies are expected to reduce substantially the land and water required¹⁶. At least in the United States the trends in agricultural technology are in the direction that should be sought in view of climatic change. Almost all technologies that are attractive for agriculture are only more attractive in light of the possibility of climatic change. Specifically, appealing directions for agricultural innovation might include diversification of crop production by varying maturity, heat and drought tolerance, input needs, and end uses; innovation in planting and spacing; collecting and recycling irrigation run-off; soil moisture conser-

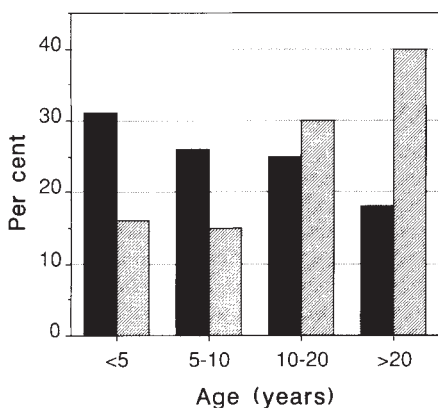


FIG. 2 Age distribution of nonmachinery capital stock in the FRG, data for 1985 (ref. 23) (hatched bars), and in the Soviet Union (solid bars), data as of 1986 (USSR State Committee on Statistics, Statistics on social indicators and capital vintage structure in industry, undated memo, Institute for Social and Economic Statistics, Moscow; courtesy of A. Grübler, Laxenburg, Austria).

vation; better moisture-use efficiency and improved use of plastics and other new materials; resistance to pests and insects; management practices; institutional measures; programmes and facilities to support extreme contingencies; and infrastructure. Many of these are applications of information technologies, as well as biotechnologies and more traditional agricultural and mechanical technologies. It may be possible to design and select plants adapted to higher concentrations of CO₂ and other changes in the atmosphere.

Though many innovations helpful in a rapidly changing climate are more likely to come from private enterprise than government, governments can help in two ways.

One way governments can aid adaptation is through timely information. One variety is assessments of the issues relating to climatic change. A second important and more operational variety of information is improved weather and climate forecasts. Eventually the climate of the far future will become tomorrow's weather. Information about it is likely to improve the possibility that it will be more resource than hazard.

There has been a gradual, measured improvement in weather forecasting during the past 20 years (ref. 17). In the greenhouse issue, all nations should find strong motivation to improve forecasts and the data and research underlying them. The quality of weather analyses and forecasts in many developing countries, especially in the tropics, is markedly lower than those in developed countries, particularly in the northern hemisphere. Advances in numerical modeling and extension of technology for monitoring in tropical regions can cause substantial improvements.

A third innovation in information relates to markets. The needs are for facilitation of information flows and improvements in rules for markets, in particular, markets for water. In many nations, water is allocated largely

through administrative means based on water rights. Water transfers accommodate new patterns of climate, as well as changing farming and urban and industrial growth. They allow water to be used where it is most valuable¹⁸. Flexibility of water transfer is important, because the life of water projects is often 50–100 years or more¹⁹.

Substantial subsidies for water for irrigation, in particular, lead to prices that encourage inefficiency. There is little incentive to conserve. Higher cost of constructing water projects and more demand and competition for water for such uses as preservation of wildlife, recreation, and cities make the issue serious. Several long-term contracts in the Central Valley of California provide water for only \$3.50 per acre-foot, whereas new sources of supply would cost the federal government or the state \$200–300 per acre-foot per year for construction alone¹⁸. Allowing and encouraging voluntary marketing of the resource among users could help adapt to climate changes and produce economic benefits. Voluntary water transfers could take several forms, including permanent sales, long-term leases, short-term leases, and leases contingent on drought.

Although markets may require innovation by government in providing information and rules, there are also traditional 'hardware' opportunities for innovation in public works. Government is the primary purchaser, financier and manager of systems of water supply and waste disposal, as well as coastal facilities. These take decades to site and construct and then can last generations. There may be opportunities for government to enhance innovation in infrastructure in light of the possibility of a changing environment.

Conclusion

Technologies are available for adaptation to climate on a spectrum of space, time and cost. Within minutes and for a few dollars one can buy an umbrella for local protection

against a shower. Such technologies diffuse rapidly into the society, in a matter of months or years. Larger and more costly innovations, like electric refrigerators, may take 20 or 30 years to become pervasive. At the other end of the spectrum, large systems, like those for water and transportation take several decades or generations to extend themselves fully and may cost tens or hundreds of billions of dollars.

Technological performance has improved throughout human history, and in this century waves of innovation have come ever more rapidly²⁰. In many systems, there have been steady improvements in efficiency of about two per cent per year, so that systems built today, for instance in energy, are about twice as efficient as those built 30 years ago²¹. Today generating a kilowatt of electricity in a steam plant takes only 15 per cent as much fossil fuel as at the turn of the century. A doubling of overall efficiency of several major systems should be possible just by replacing existing systems with best technologies and practices available today. This, of course, takes capital, and it is not clear that the expected rate of climatic change warrants an acceleration over the rate of change in physical capital stock that is already occurring, as long as the new stock is acquired with the best information about future climate in mind.

The general direction of change in technology and civilization is heartening for those anxious about climatic change. The trend is toward systems that are less vulnerable to climate. It would seem to be sensible to maintain this course and not to revert to reliance on such technologies as sailing ships and water mills that are more sensitive to climate. The highest need is probably to assure the inventive genius, economic power, and administrative competence that make the many technologies useful in adapting to climate available to the most people. □

Jesse H. Ausubel is at The Rockefeller University, New York, New York 10021-6399, USA.

1. Nordhaus, W. D. *Setting National Priorities: Policy for the Nineties* (ed. Aaron, H.) 185–211 (Brookings, Washington, 1990).
2. Adams, R. McC. & Nissen, H. *The Uruk Countryside: The Natural Setting of Urban Societies* (University of Chicago Press, Chicago, 1972).
3. Newman, L. F. (ed.) *Hunger in History: Food Shortage, Poverty, and Deprivation* (Blackwell, Oxford, 1990).
4. Desmond, K. *Harwin Chronology of Inventions, Innovations, and Discoveries from Pre-History to the Present Day* (Constable, London, 1986).
5. Schelling, T. C. in *Changing Climate* (National Research Council) 449–482 (National Academy Press, Washington, 1983).
6. Escudero, J. C. *Climate Impact Assessment* (eds Kates, R. W., Ausubel, J. H. & Berberian, M.) 251–272 (Wiley, Chichester, 1985).
7. Weihe, W. H. *Proceedings of the World Climate Conference* 313–368 (World Meteorological Organization, Geneva, 1979).
8. *New York Times*, 1 August 1990, p. C1.
9. Warrick, R. A. *Climatic Constraints and Human Activities* (eds Ausubel, J. & Biswas, A. K.) 93–123 (Pergamon, Oxford, 1980).
10. Reynolds, T. S. *Sci. Am.* 251(1) 122–130 (1984).
11. Intergovernmental Panel on Climate Change *Scientific Assessment of Climate Change, Report to IPCC from Working Group 1* (World Meteorological Organization, Geneva, 1990).
12. US Bureau of the Census *Statistical Abstract of the United States: 1990* 110th edition (US Government Printing Office, Washington, 1990).
13. Duvick, D. N. *Econ. Bot.* 38, 161–178 (1984).
14. Grübler, A. *The Rise and Fall of Infrastructures: Dynamics of Evolution and Technological Change in Transport* (Physica, Heidelberg, 1990).
15. McEvedy, C. & Jones, R. *Atlas of World Population History* (Penguin, New York, 1978).
16. Office of Technology Assessment, *Technology, Public Policy, and the Changing Structure of American Agriculture* (US Government Printing Office, Washington, 1986).
17. World Meteorological Organization (WMO) *The World Weather Watch: 25th Anniversary 1963–1988* (WMO No. 709, Geneva, 1988).
18. Wahl, R. W. *Markets for Federal Water: Subsidies, Property Rights, and the Bureau of Reclamation* (Resources for the Future, Washington, 1989).
19. Waggoner, P. E. (ed.) *Climatic Change and US Water Resources* (Wiley, New York, 1989).
20. Mensch, G. *Stalemate in Technology* (Ballinger, Cambridge, 1979).
21. Nakicenovic, N. & Grübler, A. *Technological Progress, Structural Change and Efficient Energy Use: Trends worldwide and in Austria*, International Part (International Institute for Applied Systems Analysis, Laxenburg, 1989).
22. Economic Planning Agency *Economic Survey of Japan, 1987–1988* (Tokyo, 1989).
23. Statistisches Bundesamt *Wirtschaft und Statistik* 4/89 (Metzler-Poeschl, Stuttgart, 1989).

Is Darwinism a thermodynamic necessity?

The concept of fitness has helped evolutionists avoid the accusation of tautology, but there is a case for asking that the notion should be defined in operational terms as well as by its consequences.

THE argument that the theory of natural selection is nothing but a tautology, due in its most elegant form to Sir Karl Popper, now hardly ever keeps evolutionists or population geneticists awake at night. That, no doubt, is how it should be. If artificial selection can be used to breed specialized forms of animals or plants, why should not natural selection, by the slings and arrows of the changing environment, yield species well-adapted to the world in which they find themselves?

That is a defensible position, but it does not remove the nagging doubt. Most simply, the tautology is the observation that if, in the "survival of the fittest", the fittest are identified with the survivors, the principle of natural selection amounts to the empty slogan "survival of the survivors". This proposition does not explain why there should be survivors anyway.

Evolutionists seek to avoid the difficulty by the intermediate concept of "fitness", a measure of the adaptation of an organism to the present environment that Darwin would have related to the numbers of living offspring it is likely to leave behind. There are many circumstances in which the notion works well and simply as "for example" in accounting for the survival of the sickle-cell gene in regions where malaria was once endemic; the relative immunity of heterozygotes to infection (compared with those carrying normal genes) outweighs the decrement of fitness of those carrying two copies of the gene, who are likely to suffer from congenital and often-fatal anaemia.

Latterly, fitness has been more generally defined as the likelihood that an organism will contribute genes to the future gene-pool, thus allowing for the evolution of behaviour patterns anthropomorphically called altruistic (and for the notion of the "selfish gene"), but the principle is the same. The trouble, however fitness is defined, is that it is a *post hoc* measure of the course of evolution; to make a theory of evolution, one has to enumerate all possible genetic changes, estimate the degree to which they are likely to be represented in the future gene-pool and only then to tell in what way a population is evolving. Is there not some way in which fitness in this context might be more directly defined by the degree to which the metabolic processes on which an organism depends for its survival are biochemically matched, or adapted, to the environment?

If there were such a generally applicable definition, it would be possible to prove that evolution is an unavoidable necessity or,

more accurately, to identify the environmental parameters (and their rates of change with time) conducive to evolution. One source of confusion, but occasionally of conceptual difficulty as well, is that natural selection works on individual organisms, but that what evolves is a species or even an aggregation of species.

A further difficulty is that there is an unbridged gap between the mechanisms, necessarily biochemical at root, by which an organism can be counted well-adapted to its environment and the likely selection of its genes into the next generation's gene-pool. And as things are, there are not the physical criteria by which it can be judged whether the chance of selection is maximized.

This is the position from which J.-L. Torres, a Mexican from the University of Michoacan working at the International Centre for Theoretical Physics at Trieste, embarks on an attempt to apply the rules of nonequilibrium thermodynamics to the calculation of fitness in the mechanistic sense (*Il Nuovo Cimento* **13D**, 177; 1991). His objective is to find some way of measuring an individual's fitness by the degree to which its chief metabolic processes satisfy several thermodynamic criteria.

The working criteria are straightforward enough. A high degree of fitness would imply that metabolic processes are maximally efficient and, moreover, that they are maximally productive of whatever products the organism requires of them. On the other hand, the rate of entropy production should be a minimum, as should be the rate at which energy resources are consumed. In some circumstances, other criteria — for example, the rate of biomass production — might have to be reckoned with. By listing these criteria, Torres pretends no more than that they are a collection of plausible candidates in the construction of some composite indicator of what might be called instantaneous fitness.

But does not this make a horrendously complicated problem? The task of applying criteria such as these, not necessarily compatible with each other, to all the major metabolic cycles of an organism is evidently not child's play. But Torres is merely concerned to argue that it may be a little simpler than it seems.

Using such scant experimental data as there are for the production of ATP by the conversion of glucose into water (in red blood cells) and by respiration (in animal cells), he persuades himself, for example, that the expectations of conventional equilibrium thermodynamics do not necessarily

apply. For then one would expect there to be a sharp conflict between the rate of ATP production and the efficiency of the process; the faster, the less efficient. In the classical Carnot cycle, the most efficient heat engine there can be, all processes are taken to be infinitely slow so that the power is infinitesimally different from zero. Why should ATP production be different?

In reality, Torres argues, in the nonequilibrium steady-states that prevail when ATP is produced in living organisms, physiological conditions correspond to a situation in which the measured efficiency of the reaction (measured as the fraction of free energy converted) is greater than a half and the rate of conversion is a maximum or nearly so.

In principle, at least, this points to a solution of this part of the problem. If the rate of ATP production and the efficiency of the process are to be regarded simultaneously as measures of fitness in the sense of matching the performance of an organism and the external demands of its environment, there will be a trade-off to be struck whose consequences are not nearly as depressing as page-one thermodynamics would lead one to expect.

That is a cheerful signpost, but not nearly half the real battle of defining, let alone calculating, some kind of dynamical instantaneous measure of fitness. Indeed, there is a what may be a little more than a procedural difficulty tucked away in an appendix in which Torres sets out to derive a general form for the velocity of a network of reactions catalysed by enzymes (which is a necessary step in calculating the flux of reactants and products in a steady-state). He starts from the elementary textbook statement of the law of mass action that the rate of an uncatalysed reaction is determined by the concentrations of the reactants each raised to an integral power equal to its stoichiometric coefficient, and then supposes that the effect of enzymes is to multiply each coefficient by a constant different from the number one.

Many biologists will be tempted to regard this as a pointless exercise, but they will be mistaken. What matters in, say, population genetics, may be the degree to which an organism, by the totality of its behaviour, contributes genes to the future gene-pool, but it is a laudable exercise to see whether that state of affairs is reflected in the way in which the metabolism of the organism is matched biochemically to the environment. And, for what it is worth, Torres admits that there are "nontrivial gaps in the argument", which is a seemingly sign of modesty. **John Maddox**

Running rings around supernova 1987A

Ph. Podsiadlowski

THE bright nebula around the supernova SN1987A turns out to be a true ring of material and not an all-enveloping shell as had previously been assumed. Crotts and Heathcote come to this conclusion on page 683 of this issue¹, on the basis of high-resolution spectra of the nebula.

The good news four years ago was that a supernova had occurred sufficiently close for astronomers to get the first good view in four centuries of this final, violent demise of a star. The interesting news was that, in many respects the supernova did not fit astronomers' expectations for a classical type II supernova. One of the main puzzles of the event, which is still unresolved, is the question of why the immediate supernova progenitor was a blue supergiant rather than a red supergiant, as had generally been expected for this supernova type. Crotts and Heathcote's results on the ring around SN1987A fit this confounding trend and further constrain the models for the supernova's progenitor star.

The presence of dense material around the supernova had been known since the early days of the explosion. It was first inferred from narrow emission lines in the ultraviolet spectrum². The importance of this material was immediately recognized, as it represents the last structure generated by the progenitor star and thus provides a fingerprint of its late evolution. The large ratio of nitrogen to carbon and oxygen² and the small velocity spread of this material³ indicate that it was processed by nuclear reactions deep inside the progenitor, was subsequently dredged up to the surface and ejected in a low-velocity wind when the progenitor was in its red-supergiant phase (which, in this picture, would have ended only 20,000 years before the explosion).

The surprise is that this material has the appearance of an elliptical ring^{4,5} rather than that of a spherical shell as had been assumed previously (Fig. 1). In this issue, Crotts and Heathcote¹ present high-resolution spectra of the nebula, obtained over the past two years, in which the velocity structure across the nebula is spatially resolved. By compar-

ing the line shapes of several prominent emission lines with simulated line shapes obtained from a simple dynamical model for the velocity structure of material in the nebula, they are able to demonstrate that the ring is indeed a physical ring (a circular ring

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

FIG. 1 The ring around supernova SN1987A as recorded by the Hubble Space Telescope on 23 August 1990. The ring was formed by the progenitor star before the explosion seen in 1987 and is illuminated by the supernova. The tightly knotted debris from the explosion appears as a red blob near the centre of the ring (Courtesy of NASA).

of material inclined with respect to the line of sight); the nebula's structure is therefore very different from the spheroidal shells, often seen around the central stars of planetary nebulae, which superficially have the same ring-like appearance.

Although there can be little doubt left that we are seeing a physical ring of material formed before the supernova, its origin is, at the moment, rather mysterious. The ring-like geometry of the ejecta implies that the progenitor and/or its winds had an axisymmetric, but highly nonspherical structure.

Such an asymmetry could be produced if the progenitor was a rapidly rotating supergiant. But straightforward considerations of angular momentum show that even a star which rotated rapidly during its early evolution would become a slow rotator as it expanded into a supergiant phase⁶. Although a modest amount of rotation may be sufficient to explain the elliptical appearance of many planetary nebulae⁷, it is very doubtful that it could produce the rather extreme deviation from spherical symmetry represented by a narrow ring.

On the other hand, rapid rotation could

still provide an explanation for the ring. The only ingredient that is missing is a source of angular momentum. A close binary companion could provide just such a source. It is neither implausible nor, *a priori*, unlikely that the progenitor of SN1987A was a member of a close pair of stars, as multiplicity among stars is the rule rather than the exception (the Sun being an obvious exception). In such a binary environment, there are several mechanisms that could produce a ring. The presence of a companion alone may be sufficient to gravitationally focus any mass loss

from the stars towards the pair's orbital plane⁸. Alternatively, mass transfer between the stars in an interacting binary can provide the means by which orbital angular momentum of the binary's motion is transformed into spin angular momentum of the progenitor.

Several interacting-binary models have previously been proposed (see ref. 9) to explain the unusual features of SN1987A. Most fall into one of two classes: those in which the progenitor is the result of the merger of two stars in a common envelope, and those in which the progenitor accreted a substantial amount of matter from a close binary companion. Although the details vary substantially, all the models indicate that

after the mass-accretion phase or merging process, the presupernova star would be rotating rapidly, probably have a thoroughly mixed envelope and be surrounded by a disk of material ejected in the interaction phase. In the final blue-supergiant phase (which, according to the analysis by Crotts and Heathcote¹, lasted only 20,000 years), the fast, energetic wind from the blue supergiant would sweep up the material in this disk and form the dense ring we now see. It is worth emphasizing that models of this type can provide a theoretical framework in which three of the principal unexpected features of SN1987A can be explained simultaneously: the blue colour of the progenitor; the ring surrounding it; and the chemical anomalies in the progenitor's envelope¹⁰.

Most binary models also make clear, testable predictions. In an accretion model, the companion of the progenitor has still to be present in the system and should make itself conspicuous within the next few years. In a merger model, where the progenitor would have been a single blue supergiant at the time of the explosion, confirmation could come from a detailed reconstruction of the

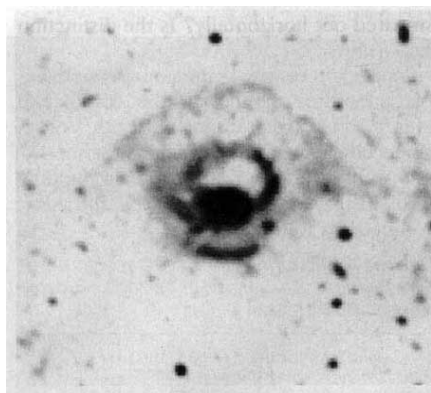


FIG. 2 The larger picture of the complex nebula around SN1987A, including the faint "Napoleon's hat". (From ref. 4.)

progenitor's history of mass loss.

In fact, the supernova ring is only part of a much larger nebula around the supernova⁴ (Fig. 2). The appearance of the whole nebula is quite complex, but clearly structured (one of its major features has the appearance of Napoleon's hat). Because the whole nebula had to be generated by the star or the stars in the neighbourhood of the supernova, it provides a direct imprint of the progenitor's eventful past, and any model, single or

binary, must be able to account for it.

As in any good mystery story, the clues are all out in the open; it is our job to decipher their meaning. This is where the next challenge in the modelling of this unusual supernova lies. While, for the moment, the ring may have added another puzzle to SN1987A, it also has provided some solid, new evidence which, in the long run, may lead us onto the right track and help to unravel this supernova's true pedigree. □

Ph. Podsiadlowski is in the Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK.

1. Crotts, A. P. S. & Heathcote, S. R. *Nature* **350**, 683–685 (1991).
2. Fransson, C. et al. *Astrophys. J.* **336**, 429–441 (1989).
3. Wampler, E. J. & Richichi, A. *Astr. Astrophys.* **217**, 31–34 (1989).
4. Wampler, E. J. et al. *Astrophys. J.* **362**, L13–L16 (1990).
5. Jacobson, P. et al. *Astrophys. J.* **369**, L63–L66 (1991).
6. Chevalier, R. A. & Soker, N. *Astrophys. J.* **341**, 867–882 (1989).
7. Kahn, F. D. & West, K. A. *Mon. Not. R. astr. Soc.* **212**, 837–850 (1985).
8. Morris, M. in *From Miras to Planetary Nebulae: Which Path for Stellar Evolution?* (eds Mennessier, M. O. & Omont, A.) 520–535 (Editions Frontiers, Gif sur Yvette, 1990).
9. Hsu, J. J. L., Joss, P. C., Podsiadlowski, Ph. & Rappaport, S. in *Supernovae* (ed. Woosley, S.) 191–197 (Springer, New York, 1990).
10. Höflich, P. *Proc. astr. Soc. Australia* **7**, 434–442 (1990).

Bermuda stretches a point

Randall M. Richardson

HOTSPOTS, identified as the deep-seated cause of certain linear chains of volcanoes that get progressively older along their length, have played an important role in the study of plate tectonics since its earliest days^{1,2}. With the Hawaiian–Emperor seamount chain as the classic and undisputed example, the number, origin and dynamics of hotspots have been hotly debated over the years. At one time, four hundred were postulated for Africa alone³. It seemed that workers correlated every large hill or linear volcanic feature with a hotspot. This trend has waned over time, as exemplified by new work by Peter Vogt⁴ providing an alternative explanation for the volcanism and elevated topography of the eastern United States and along the Bermuda Rise in the western Atlantic. He postulates two north-east trending, linear, convective upwelling zones about 1,000 km on either side of the eastern North American continental margin, separated by a downwelling zone at the continental margin. The convection is presumably locked to the continental margin by deep thermal contrasts between continental and oceanic lithosphere.

Hotspots, with sources beneath the lithosphere, have been an attractive explanation for long chains of oceanic seamounts, especially in the fast-moving Pacific plate. An idealized hotspot is pictured as a relatively narrow columnar zone, at most a few hundred kilometres in diameter, of hot mantle

material ascending rapidly through the mantle from a source in the mantle beneath the plates, if not from the core–mantle boundary itself (see figure). As a plate travels over the hotspot, a chain of volcanoes is created in sequence. Hotspots are, of course, part of the mantle convection system, which includes the plates themselves: at some point, the distinction between hotspots and other forms of mantle convection blurs. This may well be the case in Vogt's new suggestion.

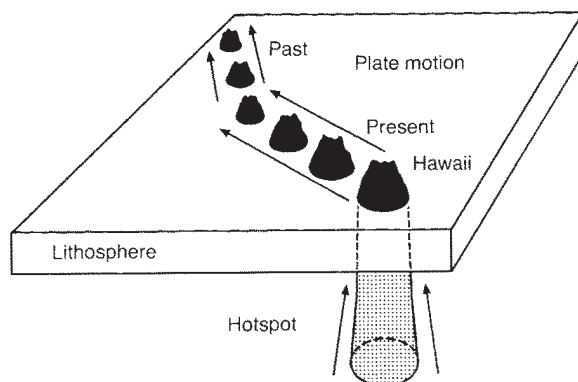
The simple hotspot model seems to work well for the Pacific. It does not work nearly as well for slow-moving oceanic areas, such as the North Atlantic. The origin of the Bermuda Rise, for example, is enigmatic. Oriented roughly parallel to the Atlantic margin, it is nearly 1,500 km long, rises 800–1,000 m above surrounding seafloor, and has other geophysical anomalies⁵ consistent with a hotspot origin. The two most difficult observations to explain in terms of hotspots are the lack of subsidence since the cessation of active volcanism 30–25 million years ago and the northeast orientation of the rise, which is nearly at right angles to the predicted motion of the North American plate.

Vogt explains these observations by hypothesizing a zone of convective upwelling paralleling the continental margin.

The hotspot model is even more difficult to apply on the continents, presumably because of the interaction between hotspot flow and continental lithosphere. The present-day topography of the Appalachian Mountains and the northeast extension of the topographic high into Newfoundland, called the Appalachian–Labrador Rise by Vogt, has been difficult to explain, with hotspots or otherwise. This average elevation of 1–2 km has persisted long after the initial construction of the mountains, even after erosion should have reduced the topography to a plain. To explain the topography, Vogt appeals to another, paired, upwelling zone on the continental side of the margin paralleling the upwelling beneath the Bermuda Rise. The upwelling zones are presumably separated by a downwelling convective zone.

Vogt's model is intriguing for a number of reasons. First, it would explain why the Bermuda Rise and the Appalachian Mountains run parallel to the continental margin. To date, a weak Bermuda hotspot, unable to burn a hole through the plate and forced to take advantage of previously existing structures, has been the best hotspot explanation for the orientation of the Bermuda Rise^{6,7}. The northeasterly trend of elevated topography on the continent has been particularly difficult to explain in terms of hotspots as well. Second, convective upwelling could also account for the long persistence, with minimal subsidence, of the topography of the Bermuda Rise and the Appalachians.

The work of Gurnis⁸ clearly indicates the important role that continental lithosphere may play in controlling convection in the mantle. In that modelling, continents tend to move toward, and stagnate above, cold downwelling zones. This inhibits subduction and may lead to continental breakup. These model results, however, do not support the existence of paired upwelling zones on either side of continental margins as proposed by Vogt. Furthermore, the return flow of material from subduction zones to mid-ocean ridges⁹ in the mantle may smear out



An idealized view of a hotspot, as a narrow column of mantle material upwelling beneath the lithosphere. Volcanoes are generated in sequence on the overriding plate as it moves over the hotspot.

Notes from underground

THE search for cosmological dark matter has taken Japanese physicists into caves in a stone quarry 100 km north of Tokyo. There, S. Orito *et al.* (*Phys. Rev. Lett.* **66**, 1951-1954; 1991) installed plastic sheets, covering 2,000 m², and left them for 2 years, shielded from cosmic rays and solar flares. Their hope was that any passing superheavy elementary particles, if such things exist, would leave in the sheets tracks of atomic damage that could be revealed later by chemical etching. The method is sensitive to slow-moving particles weighing at least 10¹² proton masses. Orito *et al.* found no sign of any superheavy particle and conclude that their abundance must be less than one for every 10²⁹ photons in the Universe — a result which pushes the cosmological limit but does not yet rule out this putative form of dark matter.

Tail tales

COMPETITION in the female reproductive tract between sperm of different males is a matter not only of sperm numbers, as is recognized, but of their length (*Proc. R. Soc. B* **243**, 181-185; 1991). M. Gomendio and E. R. S. Roldan looked at the sperm of certain primates and rodents, dividing the animals into two groups — polyandrous species, in which females mate with several males in the course of an oestrus cycle, and in which sperm competition occurs, and the singular monandrous species. The sperm of polyandrous species, it turns out, are significantly larger, by virtue of the length of their tails. The authors also report that longer sperm can indeed swim faster than their shorter brethren, and may also be able to penetrate the egg more easily — that is, size does matter.

Gender gap

R. C. GUR and colleagues, writing in *Proceedings of the National Academy of Sciences* (**88**, 2845-2849; 1991), show that men are more prone to certain age-related changes of the brain than are women. Gur *et al.* studied the crania of 69 healthy adults between the ages of 18 and 80 years using magnetic resonance imaging, and found support for the idea that ageing is, as previous work has shown, associated with a decrease in brain volume and increase in that of the cerebrospinal fluid. But they also observed that the rate of increase in volume of the cerebrospinal fluid is greater with age in men than it is in women. Moreover, men display an asymmetric effect, not seen in women, atrophy of the left hemisphere being more pronounced than that of the right. Choosing a judiciously conditional turn of phrase, the authors say that their study "may help us to understand neural substrates of behavioral changes associated with aging".

any features that might form at continental margins. A final question involves the long-term persistence of the upwelling zones. The distance of the paired upwelling from the mid-ocean ridge must change as new plate is created at the mid-ocean ridge. Thus, to persist, the upwelling must be very strongly tied to the thermal contrasts between continental and oceanic lithosphere.

The final, perhaps semantic, question is: when does a hotspot cease to be a hotspot?; or, what is the difference between hotspot convective upwelling and other forms of convective upwelling? If hotspots are limited to narrow upwelling zones moving slowly, at best, with respect to one another, then the distinction is rather straightforward. Vogt's upwelling has a long horizontal axis parallel to the continental margin and is tied to the continent rather than to other hotspots. How does this differ from a weak hotspot that gets

smear out horizontally? Is the distinction important?

Vogt's interesting argument may help explain features parallel to the continental margin, and further define hotspots. Vogt is correct in saying that one of the most important problems is the post-mountain building uplift history of the Appalachians, and that it will be hard to test his model. □

Randall M. Richardson is in the Department of Geosciences, University of Arizona, Tucson, Arizona 85721, USA.

1. Wilson, J. T. *Nature* **197**, 536-538 (1963).
2. Morgan, W. J. *Nature* **230**, 42-43 (1971).
3. Burke, K. & Wilson, J. T. *Nature* **239**, 387-390 (1972).
4. Vogt, P. R. *Geology* **19**, 41-44 (1991).
5. Detrick, R. S., Von Herzen, P. R., Sandwell, D. & Dougherty, M. J. *geophys. Res.* **91**, 3701-3723 (1986).
6. Crough, S. T. *A. Rev. Earth planet. Sci.* **11**, 165-193 (1983).
7. Liu, M. & Chase, C. G., *J. geophys. Res.* **94**, 5571-5584 (1989).
8. Gurnis, M. *Nature*, **332**, 695-699 (1988).
9. Chase, C. G. *Geophys. J. R. astr. Soc.* **56**, 1-18 (1979).

BIOPOLYMERS

Light on molecular recognition

Robert Kaiser, Tim Hunkapiller and Leroy Hood

WRITING in *Science*¹, Fodor *et al.* describe an exciting strategy for the light-directed, spatially addressable parallel synthesis of a large number of chemically related yet distinct species on a glass surface — thus producing molecular diversity through combinatorial chemical synthesis. The technique provides a fresh way of tackling one of the fundamental problems of modern biology, namely molecular recognition, or how nucleic acids and proteins interact in a highly specific manner.

Interactions between nucleic acids generally arise from one-dimensional complementarity through the ability of nucleotides to form classic Watson-Crick base pairs. Molecular recognition at the DNA level, therefore, is often assayed by determining whether a DNA probe and target sequence hybridize. In contrast, proteins interact with other molecules by three-dimensional complementarity, the matching of shapes that give rise to optimal interatomic interactions.

Because the rules for protein folding are unknown, studies of molecular recognition at the protein-protein or protein-nucleic acid levels often centre around the testing of many diverse forms of one component for specific binding to the second component. In general the process can be broken into two parts: first, the generation of numerous forms in which a limited set of elements are combined in all possible ways to generate a potentially very large and diverse set of candidates for testing; second, the selective enrichment of one or a few high-affinity forms from this test set. In the past, investigators have tended to employ biological strategies for both combinatorial diversification and selection^{2,3}. More recently chemical synthesis has been used to generate combinatorial diversification followed by biological

selection⁴. Fodor's approach provides a massively multiplexed assay system that is no longer dependent on the limitations of competitive selection.

The power of combinatorial strategies for generating molecular diversity is best illustrated by the vertebrate immune system⁵. The cells of the immune system must be able specifically to detect and bind a pool of foreign molecular species (antigens) of potentially unlimited structural diversity. The genome of a single organism cannot hope to encode a unique receptor for every possible antigen, so the limited diversity of the germline genetic information must be amplified. This is accomplished by randomly selecting and combining elements from limited pools of gene segments (V, D and J) to construct a pair of unique rearranged genes for each immune cell that encodes the immune receptors (antibodies and T-cell receptors). Thus, the potential receptor diversity of an individual organism is proportional to the product of the sizes of the germline repertoires of the various gene-segment groups⁵. It is estimated that the human immune system has the potential to generate some 10¹¹ unique antibodies by this and other combinatorial and/or mutational mechanisms⁶. Selection is then accomplished through clonal amplification of only those cells displaying a single type of receptor that complements an inducing antigen.

One particularly productive area of research that rests upon the antibody model has been the design and generation of antibody structures which mimic enzymes as selective catalysts^{2,3}. Better understanding of enzymic catalysis has allowed the generation of specific antibodies displaying a wide variety of catalytic capabilities, including hydrolysis or formation of ester and amide

bonds, acyl group transfer, β -elimination, carbon-carbon bond formation, porphyrin metalation, peroxidation and redox reactions. For example, antibody catalysts which mimic hydrolytic enzymes can be obtained using a transition-state analogue as the antigen to select for antibody-binding sites containing the appropriate amino-acid side-chain functionalities for transition-state stabilization and catalytic function. A powerful new *in vitro* combinatorial technique for generating a large repertoire of monoclonal antibodies for rapid screening for catalytic activity employs a combinatorial library of antibody recognition fragments in phage lambda⁷.

Combinatorial diversity followed by selection can also be applied to the study of almost any protein-ligand interactions in a manner directly analogous to *in vivo* antibody selection. Tuerk and Gold⁴ have developed an approach based on the selective amplification of preferred protein-binding sequences from a large pool of DNA sequences. A 110-base oligonucleotide was chemically synthesized, incorporating completely random nucleotides in an eight-base region known to interact with T4 DNA polymerase when transcribed into RNA. The resulting pool of over 65,000 discrete DNA sequence variants was then enriched for those containing high-affinity binding sequences by first transcribing the DNA templates into RNAs, treating the resulting RNA pool with a small amount of the protein, capturing protein-bound RNA species on nitrocellulose filters, reverse transcribing the bound RNA into complementary cDNA, and subjecting the cDNA pool to amplification by the polymerase chain reaction (PCR). The PCR products were then transcribed into RNA and the process was repeated for four rounds.

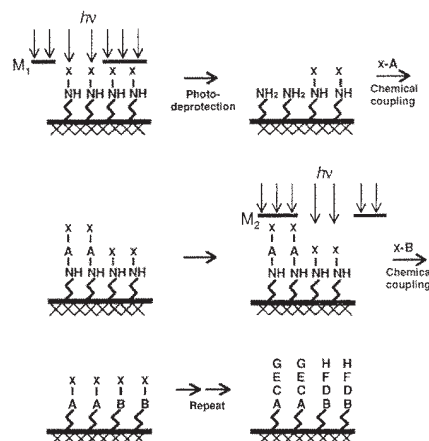
Two sequences had been selectively amplified, one with the wild-type octamer sequence, and one differing from the wild type at four unpredictable positions. Thus, diversity generated by combinatorial chemical synthesis followed by selection of high-affinity conjugates can delineate optimal binding structures. Not only does this provide a means by which to develop a deeper understanding of specific protein-ligand interactions, it should also provide insight into the fundamental problem of predicting the partial three-dimensional conformation of proteins from primary sequence information.

A difficulty with relying upon competitive selection to assay for high-affinity structures is that only the very few exhibiting optimal interactions are isolated. For some purposes, the characterization of a set of interactants exhibiting a descending degree of affinity or quality of interaction, as provided by the spatially addressable synthetic system¹, could provide a more complete picture of the interactions.

In their report¹, Fodor *et al.* describe an intriguing new approach to the chemical generation of protein or nucleic acid sequences.

The keys to the process are the elegant implementation of fairly conventional solid-phase chemistry, as well as the clever adaptation of photolithographic technology.

A glass slide is first derivatized with an appropriate silane to coat the surface with primary aliphatic amino groups, which are subsequently rendered unreactive with a photo-removable protecting group (see the figure). Light is then used to address discrete



The principles of light-directed, spatially addressable parallel chemical synthesis as described by Fodor and colleagues (reproduced from ref. 1). Amino groups attached to a glass slide are protected by a photolabile agent, x. When some groups are exposed to light through a mask (M_1) they become open to coupling to chemical building blocks (A), again each protected by a photolabile agent. The next mask (M_2) allows illumination to deprotect different amino groups, and for a different chemical building block (B) to be coupled to them. The cycle of masking, illumination and removal of the photolabile protecting agent, and then addition of further building blocks (C,D,E ...) to the exposed amino groups, continues until the products required have been synthesized.

regions of the slide by means of photo masks in a process analogous to the photolithographic techniques employed in semiconductor circuit formation. Upon illumination, only unmasked areas have the protecting group removed and are thus activated for further reaction. The patterns described by the masks used for each illumination step define the location of each synthetic product, while the molecular nature of each product is determined by which reactants are added between successive maskings. The authors demonstrate that by using simple binary masking, one of many combinatorial synthesis strategies, 2^n different polypeptide chains can be synthesized in n chemical steps using n masks. They also provide preliminary evidence for the chemical synthesis of oligonucleotides using this strategy.

Very high densities of biopolymer species can be achieved by using masks containing sufficiently small (tens of microns by tens of microns) elements. Thus, a set of thousands of different biopolymers can be synthesized on a single glass slide. Furthermore, Fodor *et*

al. show that it is possible to interrogate the various spatial elements of the array using epifluorescence microscopy in conjunction with appropriate fluorescently labelled probes. So, not only can an enormous number of variants be assayed at once, but the specific result of each individual assay can be ascertained.

An addressable system such as this has a variety of potential applications. One is in the investigation of ligand-receptor interactions, both in the exploration of basic biological processes and in the design and testing of molecules for therapeutic or diagnostic applications. Another might be the sequence analysis of DNA, using the hybridization strategy proposed by Khrapko *et al.*⁸ in which a series of fragments of template DNA are hybridized to a sorted set of all possible octamer sequences (a set with 65,256 members). It is proposed that by adjusting conditions to disallow even single nucleotide mismatches, the template sequence can be mathematically determined from the sequences of octamers forming stable hybrids.

The covalently bound biopolymer arrays thus made possible could have several advantages. The high densities obtainable provide the prospect of multiplexing tens of thousands of investigations, and it should also be possible to re-use the same array to perform several sequential assays interspersed with appropriate washes. Also, a variety of highly sensitive methods of detection (radioactivity, fluorescence, chemiluminescence, enzyme-linked reactions) could be adapted to this format fairly easily.

The method as described does suffer from some limitations, however. Repetitive yields of the peptide assembly reactions, for example, are only 85–95 per cent, thus restricting the maximum practical size of polypeptides which could be usefully generated. Additionally, the method does not appear to be viable for the preparative production of multiple biopolymers.

It is unclear from the data and descriptions provided by Fodor and colleagues just how difficult or how rapid it would be to carry out either a very large or several very small investigations into ligand-receptor interactions. However, the method they describe indeed holds great promise, which will be fully realized as powerful alternative technologies for the spatial addressing of discrete areas by light are developed. □

Robert Kaiser, Tim Hunkapiller and Leroy Hood are in the Center for Molecular Biotechnology, an NSF Science and Technology Center, California Institute of Technology, Pasadena, California 91125, USA.

1. Fodor, S. P. A. *et al.* *Science* **251**, 767–773 (1991).
2. Shokat, K. M. & Schultz, P. G. *A. Rev. Immun.* **8**, 335–363 (1990).
3. Lerner, R. A. *et al.* *Science* (in the press).
4. Tuerk, C. & Gold, L. *Science* **249**, 505–510 (1990).
5. Hunkapiller, T. & Hood, L. *Adv. Immun.* **44**, 1–63 (1989).
6. Davis, M. M. & Bjorkmann, P. J. *Nature* **334**, 395–402 (1988).
7. Huse, W. D. *et al.* *Science* **246**, 1275–1281 (1989).
8. Khrapko, K. R. *et al.* *FEBS Lett.* **256**, 118–122 (1989).

Gary J. Nabel

Bachelier and co-workers have now shown that infection by HIV stimulates binding of NF- κ B to DNA². In a promonocytic line, infection leads to a marked increase in NF- κ B binding activity and function, raising the possibility that NF- κ B has a part in perpetuating the infection and that viral genes activate NF- κ B. Several mechanisms could account for the increase in binding activity;

These findings are an elegant example of the precision of viral/host cell interactions, and raise additional questions about the molecular activation of HIV during infection and the role of NF- κ B. The HIV protease has been thought to be confined to the budding virus particle. This assumption must now be reassessed. Another possibility is that the virus particle can deliver sufficient protease to the cell to stimulate proteolysis of p105. At the same time, the possibility remains that HIV activates other cellular proteins, including protein kinases, cellular

Apparently, there is a common theme in infection by these different viruses which is not yet understood. As viral genes help to define critical steps in the regulation of cellular gene expression, these new insights may allow us to learn what the virus has already discovered. □

1. Rivière, Y., Blank, V., Kourilsky, P. & Israel, A. *Nature* **350**, 625–626 (1991).
2. Bachelierie, F., Alcami, J., Arenzana-Seisdedos, F. & Virelizier, J.-L. *Nature* **350**, 709–712 (1991).
3. Nabel, G. & Baltimore, D. *Nature* **326**, 711–713 (1987).
4. Zagury, D. *et al. Science* **231**, 850–853 (1986).
5. Harada, S. *et al. Virology* **154**, 249–258 (1986).
6. Leonard, J. *et al. J. Virol.* **63**, 4919–4924 (1989).
7. Ghosh, S. & Baltimore, D. *Nature* **344**, 678–682 (1990).
8. Poli, G. *et al. Proc. natn. Acad. Sci. U.S.A.* **87**, 782–785 (1990).
9. Gimble, J. M. *et al. J. Virol.* **62**, 4104–4112 (1988).
10. Leung, K. & Nabel, G. J. *Nature* **333**, 776–778 (1988).
11. Ballard, D. W. *et al. Science* **241**, 1652–1655 (1988).

Glasses come to order

Raymond Jeanloz and Quentin Williams

As kinetically frozen forms of liquid, glasses are characterized by a complete lack of long-range crystalline order and are the most structurally disordered types of solid known. To be sure, short-range order is typically present among nearest atomic neighbours in glasses. But the inability to resolve medium- and long-range interatomic correlations has prevented glasses and glass properties from being clarified at the microscopic level at which the crystalline state is understood¹. It is therefore a remarkable achievement that Gaskell *et al.*, who report on page 675 of this issue², have obtained the first experimental evidence for intermediate-range structural order in an oxide glass.

Glass is of widespread importance for several reasons, not the least of which are its technological applications. It is the lack of an underlying crystal structure that ultimately makes glass malleable — both mechanically, in that objects of nearly unlimited shapes can be formed, and also in the sense that electronic and other properties of glasses can be varied to best advantage for a given application. At a more basic level, there is much interest in the mechanisms by which condensed matter can become disordered. Glasses may offer significant insights into the structures and properties of melts, for example.

The archetypal glass, fused silica, is classically modelled as a continuous but disordered ('random') network of corner-linked SiO_4 tetrahedra. That is, order is considered to vanish only at intermediate to long ranges, beyond the first few coordination shells, because the tetrahedra are exceptionally strong and well defined molecular building blocks. Alkali and alkaline earth ions, which are often present in commercial glasses, are treated as 'network modifiers'. They are thought to disrupt the tetrahedral network, enhancing the disorder of the glass by counteracting the effects of the network-forming ions such as silicon.

The interplay between structure and composition has led to a rich literature in which the degree of polymerization, or the interconnectedness, of tetrahedral units have been documented for a wide variety of glasses³. Much of the evidence, necessarily indirect, comes from spectroscopic probes and has led to relatively detailed but arcane descriptions of glass structure. For example, much effort has gone into characterizing the abundances of Q^i species for $i \sim 0-4$, with Q^i being a cluster consisting of a network-forming tetrahedron corner-linked to i neighbouring tetrahedra⁴. How the Q^i speciation is related to glass or liquid-state properties is as yet unclear, but the consensus has been that at least this is a parameter that can be estimated from the observed spectra.

Now Gaskell *et al.* present new data which overturn the classical view of glass structure.

Specifically, they demonstrate that far from enhancing the disorder of the structure of CaSiO_3 glass, the network modifier, calcium, itself exhibits medium-range ordering. To show this, Gaskell and co-workers have carried out an elegant neutron-scattering experiment. By examining samples with varying isotopic contents of Ca, they obtain diffraction profiles that differ because the neutron scattering cross-section depends on isotopic composition. Appropriately subtracting the resulting diffraction profiles from each other, they are able to experimentally determine not only the pair correlation function around Ca, which is dominated by the distribution of Ca—O bond lengths, but also the Ca—Ca correlation function.

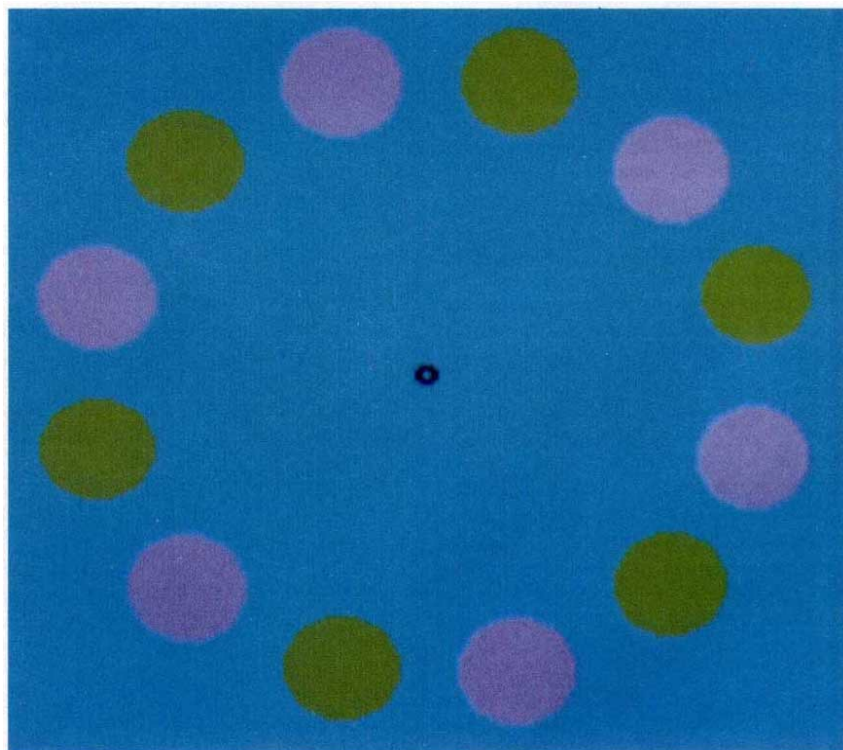
The striking feature is that two strong peaks are evident in the latter, and these can be neatly explained as being due to first- and second-neighbour Ca—Ca distances within a sheet of CaO_6 octahedra. As in the crystal, the octahedra are edge-linked. Specifically, Gaskell *et al.* show that the first Ca—Ca distance and coordination number for the glass overlap with the crystalline values; in con-

trast, a random distribution of network-modifying Ca should yield a distance about 15 per cent longer.

Through these measurements, Gaskell and colleagues identify a serious shortcoming of the classical model of oxide glass structure. Instead, they propose a model based on a nearly close-packed arrangement of oxygen ions, with the cations filling appropriate octahedral and tetrahedral interstices. This model, which gives them the sheets of edge-linked octahedra required to explain the experimental data, also has a venerable history; but for describing crystalline rather than glass structures. In brief, the new data strengthen our appreciation for the close similarity between crystalline and glass structures, at least at short to intermediate ranges.

The idea that glass and crystal structures can be closely related is supported by recent investigations of an entirely different process for making glass, pressure-induced amorphization. In this method, a crystal is compressed to pressures beyond its range of stability, the result being that it is directly converted to a glass, without melting or even heating⁵. Remarkably, many such pressure-formed glasses become crystalline again on decompression⁶⁻⁸; and in one case, we found that a crystal that was amorphized under pressure reverted, upon decompression, to a

Spot the difference



If the eyes are focused on the dot in the middle of the central black spot, after 10–15 seconds some of the peripheral coloured disks disappear and the space they occupy becomes 'filled in' by the background. This effect, the result of "an artificial perceptual scotoma [blind spot]", is described by V. S. Ramachandran and R. L. Gregory on page 699 of this issue. It is, they say, the effect supposedly used by King Charles II to 'decapitate' his courtiers, using his natural blind spot.

single crystal with the original orientation⁹. Hence, the crystalline and amorphous structures are thought to be very closely related, with the amorphization perhaps being due to a crystal-structural instability involving small displacements of the ions.

The discovery that glasses have intermediate-range order opens up the prospect of tailor-making glass structures of a given compound. Indeed, there already exists evidence that distinguishable 'polymorphs' of glass — varying mainly in medium-range order — can be reproducibly made^{10–14}. By

analogy with the polymorphism of crystalline solids, the properties of such amorphous phases should be significantly altered as the structure is changed. The potential exists, therefore, for optimizing mechanical, optical and other properties of glass by varying its newly discovered structure. □

Raymond Jeanloz is in the Department of Geology and Geophysics, University of California, Berkeley, California 94720, USA and Quentin Williams is in the Institute of Tectonics, University of California, Santa Cruz, California 95064, USA.

1. Waseda, Y. *The Structure of Non-Crystalline Materials* (McGraw-Hill, New York, 1980).
2. Gaskell, P. H., Eckersley, M. C., Barnes, A. C. & Chieux, P. *Nature* **350**, 675–677 (1991).
3. Mysen, B. O. *Structure and Properties of Silicate Melts* (Elsevier, Amsterdam, 1988).
4. Stebbins, J. F. *Nature* **330**, 465–467 (1987).
5. Hemley, R. J., Jephcoat, A. P., Mao, H. K., Ming, L. C. & Manghni, M. H. *Nature* **334**, 52–54 (1988).
6. Fujii, Y., Kowaka, M. & Onodera, A. *J. Phys. Chem.* **93**, 789–797 (1989).
7. Sankaran, H., Sikka, S. K., Sharma, S. M. & Chindambaran, R. *Phys. Rev.* **B38**, 170–173 (1988).

8. Meade, C. & Jeanloz, R. *Geophys. Res. Lett.* **17**, 1157–1160 (1990).
9. Kruger, M. B. & Jeanloz, R. *Science* **249**, 647–649 (1990).
10. Mishima, O., Calvert, J. D. & Whalley, E. *Nature* **314**, 76–78 (1985).
11. Grimsditch, M. *Phys. Rev. Lett.* **52**, 2379–2381 (1984).
12. Tse, J. S. & Klein, M. L. *Phys. Rev. Lett.* **58**, 1672–1675 (1987).
13. Hemley, R. J., Chen, L. C. & Mao, H. K. *Nature* **338**, 638–640 (1989).
14. Williams, Q., Knittle, E., Reichlin, R., Martin, S. & Jeanloz, R. *J. geophys. Res.* **95**, 21549–21563 (1990).

NEUROBIOLOGY

trking neurotrophic receptors

Cliff Ragsdale and Jim Woodgett

A MOLECULAR collision has occurred between the fields of neurotrophic factor research and protooncogene biology. On page 678 of this issue¹ and in the 5 April issue of *Cell*², the *trk* protooncogene has been identified as the elusive 140K receptor for nerve growth factor that is required for high-affinity binding.

Nerve growth factor (NGF) was the first discovered and remains the best characterized of the target-derived neurotrophic factors³. It is a survival factor for sympathetic and selected sensory neurons of the peripheral nervous system and for injured cholinergic neurons of the basal forebrain; in addition, it can regulate the outgrowth of neuronal processes and promote the expression of neural differentiation markers. The factor may have clinical relevance in the treatment of neurodegenerative disorders, so elucidation of its mechanism of action has long been a priority.

The molecular cloning of a receptor of relative molecular mass 75,000 (*M_r* 75K) for NGF^{4,5} heralded a breakthrough, but excitement was quashed by the finding that this protein exhibited only low-affinity binding to the factor, some 100-fold lower than the dissociation constant *K_d* of 10⁻¹¹ M linked to such biological effects as neurite outgrowth. Even so, from gene-transfer experiments it seemed that this low-affinity receptor (LNGFR) participated in the high-affinity binding. The search for the 'high-affinity' receptor was bolstered by work from Shooter's laboratory that identified a second NGF-binding protein of 140K in cells displaying high-affinity binding sites. Furthermore, this protein contained phosphotyrosine⁶, implying that it might participate in a tyrosine kinase signalling pathway.

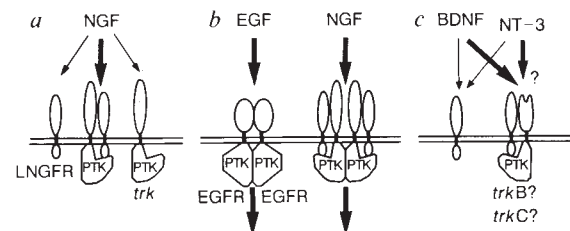
Meanwhile, interest in the *trk* family of receptor-like protein-tyrosine kinases was stimulated by their apparent confinement to tissues of neural origin^{7,8}. The *trk* oncogene was originally identified by transfection of DNA from a human colon carcinoma⁹: the oncogene consists of a receptor-like protein-tyrosine kinase domain fused to tropomyosin sequences (hence *trk*, tropomyosin receptor kinase). There are at least three cellular family members, *trk*, *trkB* and *trkC*, which are 130–140K proteins each harbouring an extracellular domain and an intracellular tyrosine kinase domain. Despite their receptor-like structure their putative ligands evaded identification. A flurry of papers from three groups now shows that the *trk* ligand is NGF: *trk* kinase activity in PC12 cells is activated minutes after the addition of NGF¹⁰; p140*trk* can be chemically cross-linked to NGF and is immunologically

indistinguishable from the native 140K NGF-binding protein^{2,11}; cellular expression of *trk* can confer high-affinity binding for NGF^{1,2}; and microinjection of *trk* RNA into *Xenopus* oocytes bestows an NGF-dependent meiotic maturation¹².

So what is the role of the 75K LNGFR? NGF-binding data from the groups of Chao and Parada suggest that, in isolation, *trk* binds NGF with a low affinity similar to that of the 75K LNGFR¹¹ and that high-affinity binding requires expression of both proteins in the same cellular membrane¹. Moreover, the 75K protein is clearly implicated in NGF action. In addition to the gene-transfer experiments¹, domain-swapping between the EGF receptor and the LNGFR has demonstrated a role for the transmembrane and cytoplasmic regions of the 75K protein in mediating signal transduction¹³. In this scheme the LNGFR and *trk* proteins are proposed to form a heterodimer that conveys the biological signal (see *a* in the figure). Barbacid's laboratory, by contrast, have presented evidence that *trk* expression alone is sufficient to form high-affinity NGF-binding sites². In addition, when *trk* is introduced into oocytes, a biological response to NGF is conferred, albeit an unnatural one¹².

Which model is correct? Receptor protein-tyrosine kinases, such as the prototypic epidermal growth factor (EGF) receptor, are known to dimerize for signal transmission (see *b* in the figure). Hence, high levels of expression of *trk* may, in the presence of NGF, allow some *trk* homodimer formation and thus generate high-affinity binding sites. Interestingly, the affinity of NGF for *trk* alone, as reported by Barbacid's group, does not appear to be as high as that found for the heterodimer. Given recent evidence suggesting the existence of distinguishable 'high-affinity' binding sites for NGF¹⁴, it will be important to determine the molecular combinations of these receptor subunits required for biologically important responses to NGF such as neurite outgrowth.

Just as *trk* is a member of a family of related protein-tyrosine kinase receptors, so NGF is one of at least three related neurotrophic factors, the two others being brain-



In *a*, nerve growth factor (NGF) binds with low affinity to the monomeric p75 NGF receptor (LNGFR) and to *trk*. High-affinity binding is found when both molecules are present¹. In *b*, the biologically active form of the epidermal growth factor receptor (EGFR) occurs upon dimerization of the protein-tyrosine kinase domains (PTK). Is this the case for the high-affinity NGF receptor? In *c*, two neurotrophic factors related to NGF, brain-derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3)¹⁵, may act through other members of the *trk* PTK family such as *trkB*.

derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3). The three share structural similarities, but serve as neurotrophins for largely non-overlapping populations of neurons in the central and peripheral nervous systems¹⁵. An attractive possibility is that the different *trk* proteins may serve as p140 receptor subunits for the different neurotrophins (see *c* in the figure). Interestingly, BDNF and NGF bind to the 75K LNGFR with similar affinities¹⁶, so it is possible that the LNGFR may act as a common receptor subunit for the neurotrophins. Finally, given

the existence of truncated forms of *trkB* that contain the ligand-binding domain but not the kinase domain⁸, it may be that there is a combinatorial heterogeneity to this system, in that the neurotrophin receptors may interact or cross-compete for soluble factors.

'Nerve growth factor' has long been considered something of a misnomer because NGF seems to do nearly everything except stimulate neurons to divide. One striking aspect of the current work is the finding that NGF actually works through a canonical growth-factor receptor, that is, a protein-tyrosine kinase receptor. Moreover, this

receptor has oncogenic potential when deregulated. Apparently, the pathways used in signalling cell division in mitogenic cells can be exploited to quite different effect in post-mitotic neurons. The identification of *trk* as a component of the biologically relevant NGF receptor should guarantee rapid progress as the lessons learnt from study of the mitogenic receptor kinases are applied to NGF biology. □

Cliff Ragsdale and Jim Woodgett are at the Ludwig Institute for Cancer Research, 91 Riding House Street, London W1P 8BT, UK.

1. Hempstead, B. L., Martin-Zanca, D., Kaplan, D. R. & Chao, M. V. *Nature* **350**, 678-683 (1991).
2. Klein, R. et al. *Cell* **65**, 189-197 (1991).
3. Levi-Montalcini, R. & Angeletti, P. U. *Physiol. Rev.* **48**, 534-569 (1968).
4. Johnson, D. et al. *Cell* **47**, 545-554 (1986).
5. Radeke, M. J. et al. *Nature* **325**, 593-597 (1987).
6. Meakin, S. O. & Shooter, E. M. *Neuron* **6**, 153-163 (1991).
7. Martin-Zanca, D., Barbacid, M. & Parada, L. F. *Genes Dev.* **4**, 683-694 (1990).
8. Klein, R. et al. *Cell* **61**, 647-656 (1990).
9. Martin-Zanca, D., Hughes, S. H. & Barbacid, M. *Nature*

- 319**, 743-748 (1986).
10. Kaplan, D. R., Martin-Zanca, D. & Parada, L. F. *Nature* **350**, 158-160 (1991).
11. Kaplan, D. R. et al. *Science* **252**, 554-558 (1991).
12. Nebreda, A. R. et al. *Science* **252**, 558-561 (1991).
13. Yan, H., Schlessinger, J. & Chao, M. V. *Science* **252**, 561-563 (1991).
14. Weskamp, G. & Reichardt, L. F. *Neuron* **6**, 649-663 (1991).
15. Thoenen, H. *Trends Neurosci.* **14**, 165-170 (1991).
16. Rodriguez-Tebar, A., Dechant, G. & Barde, Y.-A. *Neuron* **4**, 487-492 (1990).

FUSION ENERGY

The moment of truth nears

Richard D. Petrasso

THE generation of copious, energetic, charged fusion products (α -particles) and the deposition of their energy within the fusing plasma are pivotal issues for the achievement of fusion energy. With less than nearly perfect α energy input, the plasma cannot sustain the high temperatures (250 million K) needed to maintain deuterium-tritium fusion reactions. In preparation for the first ever experiments on deuterium-tritium plasmas in July 1993 — to occur in the US Tokamak Fusion Test Reactor (TFTR) — a workshop* was held last month to debate the critical issues regarding the plasma physics of α -particles and tritium operations.

Since the first fusion experiments in the early 1950s — conducted mainly with devices called pinches, mirrors and stellarators — all magnetically confined plasmas have been principally composed of weakly fusing, nonradioactive hydrogen isotopes, either hydrogen (mass 1) or deuterium (mass 2). But it was understood that with the eventual attainment of high-quality fusion plasmas, experiments would switch to a mixture of deuterium and tritium (mass 3), the latter being a radioactive β -emitter. Despite the special care required in handling tritium, the overriding reasons for its choice are, first, that the deuterium(d)-tritium(t) reaction



has by far the largest reaction rate and, second, that it occurs at the lowest temperature (n is an energetic neutron, which in principle would be used to power conventional generators, and α is an energetic He^{2+} nucleus).

The large kinetic energy of the α and neu-

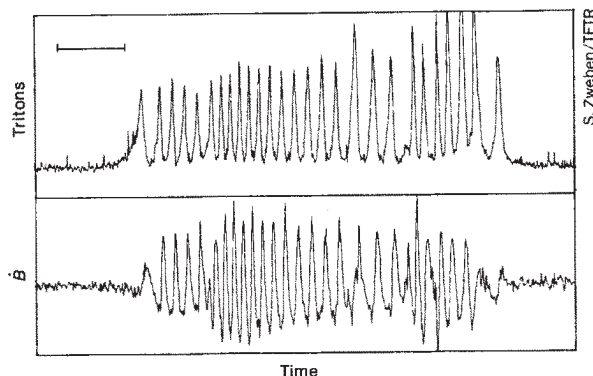
tron results from the strong nuclear interaction which converts 0.4 per cent of the deuterium and tritium mass into energy. In contrast ordinary chemical reactions are a million times less energetic, though they too derive their energy by converting mass (as in $mc^2 = E$). But it is precisely because of this million-fold enhancement that, on the one hand, the frightful power of the hydrogen bomb has been realized and, on the other, two generations of physicists and engineers have pursued the goal of fusion energy. One of the principal results of this effort has been the tokamak, a toroidal plasma-containment device in which a large magnetic field (about 50 kilogauss) and plasma current (around 3 megamp) thread the torus, thereby inhibiting the loss of energetic plasma particles.

A figure of merit that represents the quality of any fusion plasma, be it the Sun's core or a tokamak, is $Tn\tau$, where T is the central ion temperature, n the central ion density, and τ the overall energy confinement time (the duration for which the plasma thermal energy can be effectively contained). For a deuterium-tritium plasma to ignite — that is, for the α -energy input to sustain the $250 \times 10^6 \text{ K}$; $n = 10^{20} \text{ cm}^{-3}$; and $\tau = 3 \times 10^{-4} \text{ s} = 10 \text{ million years}$). The fusion plasma experiments

of the 1950s were a factor of 10^7 below this ignition threshold. In the past few years TFTR¹ and JET² (Joint European Torus) obtained plasma conditions which, had they been operating with deuterium-tritium under identical conditions, would have momentarily achieved a $Tn\tau$ factor of 20 and 6 below this threshold. (For JET, $T = 250 \times 10^6 \text{ K}$; $n = 3.7 \times 10^{13} \text{ cm}^{-3}$; and $\tau = 1 \text{ s}$.) Thus although neither tokamak will achieve ignition — nor were they designed to — each can be expected to produce measurable amounts of α power (2 megawatts) and α -particles ($3 \times 10^{18} \text{ s}^{-1}$) if using deuterium-tritium fuel. But most of all, several critical issues can be experimentally addressed for the first time.

One such issue is the energy and particle transport in a deuterium-tritium plasma (D. Post, TFTR). Although one might not expect that changing from one hydrogen isotope to another would notably affect the transport, experience teaches otherwise. For example, plasmas of pure deuterium typically run hotter than those of pure hydrogen, and this is partly reflected in a larger value of τ . A corollary is, however, that impurities are also better confined in a deuterium plasma³; if inadvertently injected in too large quantities, these could radiatively quench the plasma, as occurred on the record JET discharge^{2,4}. Thus it will be important to investigate whether the empirical scaling for the energy confinement, $\tau \sim m_i^{1/2} / P^{1/2}$, still holds (m_i is the ion mass and P the total absorbed power), and whether impurities are more likely to reside within the plasma precluding a significant and sustained value of $Tn\tau$.

A second crucial question is what fraction of α -particles will be either kicked out by intrinsic plasma instabilities — such as by sawtooth oscillations and higher-frequency magnetic perturbations (S. Zweben, TFTR; see figure) — or execute a process known as stochastic diffusion whereby they make a rapid random-walk out of the plasma before depositing their energy in the plasma, which takes 0.3 s according to classical calculations



Escaping 1-MeV tritons (upper) resulting from magnetic perturbations δB (lower) associated with magnetohydrodynamic modes. The energetic tritons are detected by particle detectors just outside the plasma boundary. An important issue is whether these perturbations or others, such as those associated with the α -driven Alfvén mode, will eject a large fraction of the energetic α -particles generated in TFTR deuterium-tritium plasmas. (Scale bar, 1 ms.)

*The 1st Workshop on Alpha Physics in the Tokamak Fusion Test Reactor, Princeton, 28-29 March 1991.

(R. Boivin, TFTR). In principle, each of these effects has been observed at TFTR through energetic protons (p) and tritons (t) created in the fusion reaction



And in closely related work at JET (F. Marcus), 14.7-MeV protons from deuterium-helium-3 reactions were poorly confined with a 1.5 MA plasma current, but well confined with a 4.5 MA current unless sawtooth oscillations occurred. For deuterium-tritium plasmas, the crux of the issue will be quantitative: what fraction of the α -particles will be so affected, and how often?

A conceptually related issue is whether the large population of α -particles will themselves generate and drive formerly benign or unknown collective oscillations in the plasma that would again result in their 'self ejection'. Of the class of such instabilities that physicists theorize, none was more widely discussed than the α -driven Alfvén mode⁵⁻⁷. This instability should occur if a sufficient number of α -particles have, along the direction of the magnetic field, a component of velocity greater than a critical velocity called the Alfvén velocity, V_A (D. Sigmar, MIT). (The tokamak's confining magnetic field includes a toroidal component and a smaller poloidal field created by the plasma current.)

The Alfvén velocity, which depends on the field strength (B) and central ion density ($V_A = B/\sqrt{4\pi m_i n_i}$), is around $0.8 \times 10^9 \text{ cm s}^{-1}$ (with a field of 50 kilogauss and central ion density of $6 \times 10^{13} \text{ cm}^{-3}$). The initial velocity of the 3.5-MeV α -particles is $1.3 \times 10^9 \text{ cm s}^{-1}$, and about 40 per cent will have a 'parallel' velocity component exceeding the Alfvén velocity.

But it is also becoming apparent that two other conditions need to be satisfied if the Alfvén mode is to be excited. First, the number density of α -particles at the centre must not be less than about 0.3 per cent of the electron density. And second, the α -particles' gas-kinetic pressure must exceed 0.003 times the magnetic pressure they experience ($B^2/8\pi$, about 100 atmospheres).

Simulations of TFTR deuterium-tritium plasmas (R. Bundy) find all three thresholds likely to be exceeded. But recent " α -simulation experiments" (K. Wong) on TFTR using energetic neutral beams and deuterium plasmas indicate that the pressure threshold may be larger by about an order of magnitude⁷. If so, then TFTR deuterium-tritium plasmas may be spared this particular α -particle loss channel.

The best method of injecting tritium into TFTR is also of pressing importance. The most frequently discussed methods are injecting it either with high-power heating beams, or firing in tiny frozen tritium pellets. Among other compelling arguments for using pellets (each less than 0.01 g in mass, equivalent to 100 curies of radioactivity), one is that it will tend to place the tritium mostly on the tokamak axis where it can be

more effectively burned (D. Mikkelsen, TFTR). Another not inconsequential benefit is that pellet injection reduces by an order of magnitude the total amount of tritium injected into the machine (D. Meade, TFTR).

Making maximum use of the limited number of deuterium-tritium discharges is another concern (J. Strachan, TFTR); only 1,000 or so are feasible, whereas TFTR has already had about 50,000 hydrogen, deuterium and helium discharges. For example, should new diagnostic methods be developed that in principle could, with some spatial and temporal resolution, measure aspects of the α -particles' energy distribution? Although highly desirable, the means to measure these parameters are at present untested, although serious efforts to do just that are underway (P. Woskov, MIT). And even if not immediately applicable to TFTR, they might prove invaluable for the next generation of experiments.

An approach that addresses the immediate needs of TFTR is to place greater emphasis on proven methods that could measure the escaping α -particles (as with tritons as shown in the figure). Largely in combination with the α -particle birth profile (from 14.1-MeV neutron imaging), such measurements should reveal a clear picture of the overall α -particle transport.

Furthermore, it will be essential to establish experimentally for each diagnostic technique — such as those used to measure the ion-temperature profile or the plasma impurity content — its ability to operate reliably in the intense neutron and γ -ray fluxes resulting from the fusion neutrons. None of the tokamak diagnostic techniques has yet operated in such a severe environment, so that there is concern that, despite careful shielding and other precautions, some measurements might give false or unintelligible signals owing to the background radiations.

Finally, assuming that most of the α -particles do deposit their energy in the plasma, what is to be done with the thermalized He^{2+} 'ash' which would certainly dilute the deuterium-tritium fuel and so extinguish high-temperature ignition? One proposed solution (G. Miley, Univ. Illinois) is that suitable plasma instabilities should be periodically triggered so as to expel the ash. However, from this and previous considerations it is clear that a fine line has to be walked: on the one hand energetic α -particles must deposit their energy in the plasma; on the other, having done so, they must be transported out. TFTR should give us our very first glimpse of this delicate balance. □

Richard D. Petrasso is at the Plasma Fusion Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Dead Heat

LAST week Daedalus proposed a novel sound-absorber. It is a rubber plasticized with high-pressure gas, so that the contractile tension in the rubber matrix is exactly balanced by the pressure of the occluded gas. Over a wide density-range it has no preferred volume and zero elasticity. So sound traverses it with zero velocity, and never gets anywhere.

Both as a perfect sound-insulator, and as a completely 'dead' vibration-damper, the new polymer/gas composite will find innumerable applications. But Daedalus now points out that heat is also a form of sound; it is propagated by thermal phonons. A gas-swollen polymer balanced to have zero phonon-velocity should be a perfect heat-insulator.

At last the bulkiness and inconvenience of conventional lagging will be overcome. Even in thin sheet form, DREADCO's new "Dead Heat" polymer will totally insulate refrigerators and ovens, hot-water tanks and pipes, walls and windows, floors and ceilings, and all other surfaces through which heat leaks wastefully away. Offices will be fully heated by their content of staff and computers, frozen foods and hot drinks will be sold on open shelves in Dead Heat cartons, and a new market may open up for well-wrapped parcels of industrial waste heat. And on the human scale, the unstated modern tension between fashion and domestic heating will at last be resolved. Respectable victorians, with many layers of woollen underwear, could tolerate houses at 10°C . We demand a wasteful 20°C in order to show off our bodily elegance under a minimum thickness of tight-fitting clothing. But tailored in DREADCO's Dead Heat fabrics, even the thinnest jeans, most figure-hugging gowns or negligible negligees will be as warm and comfortable as arctic survival-gear.

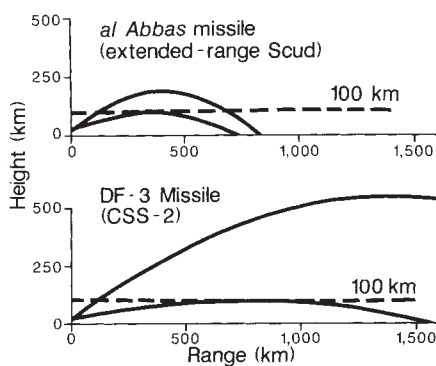
They may feel a bit like it, too. The new fabrics, with their complete absorption of movement and vibration, will insulate the wearer from the innumerable small touches and contacts which make up kinetic awareness. The resulting sense of isolation and numbness could be quite disturbing, especially in clothes like gloves, boots and underpants. The material may have to be slightly unbalanced, giving it a slight vibration-velocity.

Such a material would also make an intriguing thermal delay-line. A high temperature applied to one side of it would emerge from the other side maybe hours later. This could be very useful. A housewall of the product, for example, could be formulated to have a twelve-hour delay. More economical even than a total thermal insulator, it would retard the warmth of the day till night-time, while allowing the cool of midnight into the premises around noon. David Jones

1. Meade, D. *et al.* IAEA 1990, W. D. C. (Proc. to be published).
2. Keilhacker, M. *et al.* *Phys. Fluids* **2**, 1291 (1990).
3. Marmar, E. *et al.* *Nucl. Fus.* **22**, 1567 (1982).
4. Petrasso, R. D. *Nature* **343**, 21–22 (1990).
5. Fu, G. Y. & Van Dam, J. *Phys. Fluids* **B1**, 1949 (1989).
6. Sigmar, D. S., Hsu, C., White, R. & Cheng, C. Z. MIT Plasma Fusion Center, Rep. No. PFC/JA-89-58 (1989).
7. Wong, K. L. *et al.* *Phys. Rev. Lett.* **66**, 1874–1877 (1991).

Underflying Brilliant Pebbles

SIR — Following the use of the Patriot missile in the Gulf War, the US Strategic Defense Initiative Office (SDIO) claims that its proposed space-based antimissile system Brilliant Pebbles could be used to destroy theatre ballistic missiles such as the extended-range Scud (the *al Abbas*), but its claim ignores a simple and effective countermeasure. SDIO has acknowledged that Brilliant Pebbles can intercept missiles only above about 100 km. Accordingly, if a missile's trajectory is adjusted so that it stays below this altitude throughout its flight, the missile becomes immune to Brilliant Pebbles. Our calculations show that flying theatre missiles on such depressed trajectories is not technically demanding; missiles



Minimum-energy maximum-range trajectories, and depressed trajectories with apogees of 100 km for the *al Abbas* and DF-3 missiles. The *al Abbas* has a maximum range of 830 km with a 250-kg warhead, and reaches an apogee of 190 km; on the depressed trajectory, it still has a range of 720 km. With a 2,000-kg payload, the DF-3 has a maximum range of 2,780 km with an apogee of 550 km; on the depressed trajectory, its range would be 1,550 km.

with ranges considerably greater than those used by Iraq can underfly Brilliant Pebbles on trajectories with apogees of 100 km or less.

The Brilliant Pebbles altitude limit results from the high atmospheric densities at low altitudes; flying through regions of dense atmosphere at high speeds would blind the interceptors by heating their sensors. On 12 February 1991, SDIO director Henry Cooper stated that Brilliant Pebbles could not attack targets below roughly 60 miles (100 km). Lowering this altitude limit is difficult because of the exponential increase in atmospheric density with decreasing altitude. For example, in going from an altitude of 100 to 80 km, the density and resulting heating increase by a factor of 40.

To determine whether a missile could be flown on a depressed trajectory, we considered whether doing so would lead to unacceptable increases of stress ('loading') on the missile or heating on the re-entry vehicle and how it would affect the range and accuracy. We calculated the missile trajectories by numerically integrating the equations of

motion¹, including forces due to gravity and atmospheric drag (lift was ignored). We compared the loading and heating for the depressed trajectory to that of the maximum-range minimum-energy trajectory for the same payload. The loading is roughly proportional to ρv^2 , where ρ is the atmospheric density and v is the missile velocity; for the coefficient, we used measured values for the V2 missile². The aerodynamic heating on re-entry is roughly proportional to ρv^3 , and was calculated using equations for laminar and turbulent hypersonic flow (assuming a boundary layer transition at 30–35 km)³.

To be specific, we considered two theatre-range missiles deployed in the Middle East: the 800-km Iraqi *al Abbas* and the Chinese-built 2,800-km DF-3 (CSS-2) missile deployed by Saudi Arabia. For the *al Abbas*, we assumed a total mass of 6,900 kg, fuel fraction of 80%, payload mass of 250 kg, specific impulse of 245 s, burntime of 115 s and booster diameter of 0.9 m (refs 4, 5 and S. Fetter, personal communication). For the DF-3, these parameters were taken to be 65,500 kg, 93.7%, 2,000 kg, 241 s, 142 s and 2.25 m (ref. 6). The weight-to-drag ratio for the DF-3 re-entry vehicle was taken as 48,000 N m⁻².

Calculations were done for depressed trajectories with apogees of 100 and 80 km. The loading was found to increase by 5–7% for the *al Abbas*, and 10–11% for the DF-3, which should be within the tolerances of both missiles. The heating was found to increase by 25–30% for the *al Abbas*, and 20–25% for the DF-3. This increase should not present a problem for the relatively crude *al Abbas*, for which the entire missile acts as the re-entry vehicle. If the increase in heat absorbed by the DF-3 RV exceeds its design tolerances, the heat shielding could be increased slightly, or the heating could be reduced to that of the standard trajectory by reducing the payload by 25%.

Flying missiles on depressed trajectories decreases their range; the percentage change increases with range (see figure). However, missiles with ranges of less than 1,500 km are sufficient to threaten most targets in the Middle East.

Our estimates suggest that the accuracy of these theatre ballistic missiles, which are very inaccurate, would not change significantly. Although errors due to atmospheric effects would increase on low-altitude trajectories, the dominant errors for these missiles are caused by the guidance system and would not be substantially affected. The inaccuracy of the *al Abbas*, estimated at 3–5 km (ref. 7), is not expected to change. The DF-3 inaccuracy of roughly 2.4 km (ref. 8) might increase by a factor of 2 to 3. As the original inaccuracy is so large that the only targets would be cities, such an increase would not remove its utility as a terror weapon.

Thus, theatre missiles like the *al Abbas*

and DF-3 missiles would be able to fly on low-apogee depressed trajectories, thereby rendering them invulnerable to space-based defences.

DAVID C. WRIGHT

Federation of American Scientists,
307 Massachusetts Avenue, NE,
Washington DC 20002, USA

LISBETH GRONLUND

Center for International Security
Studies,
University of Maryland,
7100 Baltimore Boulevard,
Suite 400, College Park,
Maryland 20740, USA

1. Regan F. J. *Reentry Vehicle Dynamics* 287 (American Institute of Aeronautics and Astronautics, New York, 1984).
2. Sutton, G. P. *Rocket Propulsion Elements* (3rd ed) 199 (Wiley, New York, 1965).
3. Anderson, J. D. *Hypersonic and High Temperature Gas Dynamics* 288–291 (McGraw-Hill, New York, 1989).
4. Zaloga, S. *International Defense Review* **21**, 1423 (1988).
5. Cochran, T. B., Arkin, W. A., Norris, R. S. & Sands, J. I. *Soviet Nuclear Weapons* 220, (Ballinger, Harper and Row, Cambridge, Massachusetts, 1989).
6. Huang, Z. & Ren, X. *Space Technology* **8**, 371 (1988).
7. *Jane's Defence Weekly* **15**, 103 (1991).
8. Shuey, R. D. et al. *Missile Proliferation: Survey of Emerging Missile Forces* 65 (Congressional Research Service, Washington, 1988).

Is C₆₀ stiffer than diamond?

SIR — The availability of single crystals of C₆₀ (ref. 1) should allow measurements of the stiffness of these molecules. By using a simple elasticity argument, we estimate the bulk modulus of individual C₆₀ molecules to be 843 GPa, which is greater than that of diamond (441 GPa). We expect that the modulus of single crystals of C₆₀ will reach 642 GPa at modest pressures and that crystalline C₆₀ will then be stiffer than diamond at the same pressure.

Consider a single crystal of graphite with equal orthogonal tensile (or compressive) stress components applied to areas normal to the basal plane. Let x_3 be the direction along the c axis and x_1 and x_2 lie in the basal plane. The stress state is

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ 0 \end{bmatrix}$$

Hence the longitudinal strains ϵ_1 and ϵ_2 are given by

$$\epsilon_1 = \epsilon_2 = S_{11}\sigma_1 + S_{12}\sigma_2 = (S_{11} + S_{12})\sigma$$

where the S_{ij} are elastic compliances. The dilation ΔA of the basal area A is

$$\frac{\Delta A}{A} = \epsilon_1 + \epsilon_2 = 2(S_{11} + S_{12})\sigma$$

If h is the height of a graphite basal plane (one half of the lattice spacing c because of the ABAB stacking sequence) then

$$\frac{\Delta A}{A}h = 2(S_{11} + S_{12})\sigma h = 2(S_{11} + S_{12})\gamma \quad (1)$$

where γ is the membrane tension for a single graphite layer. (This behaves as a surface tension). If we now assume that γ is curvature-independent and that the molecules are spherical with radius R , the hydrostatic tension P_t is given by the Laplace equation $P_t = 2\gamma/R$. If we treat the C_{60} molecule more correctly as a truncated icosahedron rather than a sphere, the right-hand side of this equation must be multiplied by 1.0881. Substituting for γ in equation (1),

$$P_t = \frac{h}{R} \frac{1.088}{(S_{11} + S_{12})} \frac{\Delta A}{A} \quad (2)$$

or, in terms of the volume V ,

$$P_t = \left[\frac{2}{3} \frac{h}{R} \frac{1.088}{(S_{11} + S_{12})} \right] \frac{\Delta V}{V} \quad (3)$$

and so we obtain for the bulk modulus at zero pressure,

$$B_0 = \frac{2}{3} \frac{1.088}{(S_{11} + S_{12})} \frac{h}{R} \quad (4)$$

As this argument is based on linear elasticity, hydrostatic pressure (instead of tension) would give the same result. As the molecule is treated here as an elastic continuum, under hydrostatic pressure the pressure everywhere within the molecule is the same: the pressure at the molecular radius R_{AC} is the same as the external pressure P . Hence we do the energy balance leading to equation (4) at the surface of the truncated icosahedron defined by R_{AC} . For a quantitative estimate of B_0 we use the experimental quantities² $S_{11} = 0.00098 \text{ GPa}^{-1}$, $S_{12} = -0.00016 \text{ GPa}^{-1}$ and $h = c/2 = 3.354 \text{ \AA}$ (ref. 3). For the C_{60} ball, $R_{AC} = 3.52 \text{ \AA}$ (this is the distance from the centre of the ball to the centre of the atoms on the ball's surface (T. Siegrist, personal communication)). This yields $B_0 = 843 \text{ GPa}$, which is larger than the modulus of 441 GPa found experimentally for diamond⁴.

The molecular radius obtained from the lattice parameter of the closest-packed f.c.c. crystal is $R_{CP} = 5.02 \text{ \AA}$ (ref. 1). When C_{60} balls are placed on a f.c.c. lattice, they interact essentially via van der Waals attractive forces (as do Ne, Ar and Kr) and initially the crystal would be compliant (the bulk modulus would be relatively small). When the pressure is increased to the point where the hard spheres touch, however, the bulk modulus of the crystal will become similar to that for a single molecule: the filling factor for balls in an f.c.c. crystal is 0.74, so that using the volume-fraction rule we obtain $B = 843 \times 0.74 = 624 \text{ GPa}$ for such

(presumably modest) pressures.

The hardness of materials is related to their modulus⁵. It is thus possible that at pressures of about 20 GPa or more (where the soft-sphere repulsion is overcome), f.c.c. C_{60} crystals will be harder than diamond, and even harder than the hypothetical C_3N_4 compound suggested by Cohen⁶.

RODNEY S. RUOFF

IBM Research Division,
T. J. Watson Research Center,
Yorktown Heights,
New York 10598, USA

ARTHUR L. RUOFF

Department of Materials Science
and Engineering,
Cornell University,
Ithaca, New York 14853, USA

Plastid genes and parasitic plants

SIR — Howe and Smith¹ have provided one explanation for the retention of a plastid genome in achlorophyllous *Epifagus virginiana*, as reported by dePamphilis and Palmer². But porphyrin biosynthesis is not the only significant metabolic activity restricted to plastids; the biosynthetic pathways for glutamate, lysine, threonine, methionine, isoleucine, leucine, valine, tryptophan, phenylalanine, tyrosine, arginine, cysteine, serine and glycine are wholly or partly located within the plastid³. Thus the synthesis of most of the protein amino acids, and all the metabolites derived from them, is dependent on intact and functional plastids. Other metabolic pathways, not necessarily linked to photosynthesis, are also localized in plastids.

Even in a plant that has lost one aspect of plastid metabolism, namely photosynthetic carbon fixation, there would be little advantage in losing all plastid functions. The genes for all the biosynthetic enzymes for the pathways mentioned are encoded in the nucleus, and the polypeptides produced from these genes have leader sequences directing them to plastids. It is thus hard to envisage any evolutionary process in which all these enzymes could be redirected (simultaneously?) to other parts of the cell, and their functions integrated there.

The plastid is an essential component of the plant cell — the chloroplast is merely a specialized plastid restricted to a few cell types and tissues. So the retention of a plastid genome in *E. virginiana* is not remarkable, although its absence must surely be.

ROGER M. WALLSGROVE

Soils and Crop Sciences Division,
Rothamsted Experimental Station,
Harpenden,
Hertfordshire AL5 2JQ, UK

1. Howe, C. J. & Smith, A.G. *Nature* **349**, 109 (1991).

2. dePamphilis, C. W. & Palmer, J. D. *Nature* **348**, 337–339 (1990).

3. Lea, P. J., Wallsgrove, R. M. & Miflin, B. J. in *Chemistry and Biochemistry of the Amino Acids* (ed. Barrett, G.C.) 197–226 (Chapman & Hall, London & New York, 1985).

Muscle damage in mdx mice

SIR — Menke and Jockusch reported¹ that hypo-osmotic shock exposes a reduced stability of dystrophin-deficient mdx mouse muscle. However, this is not apparent with experimental models of damage to mdx mouse muscle induced by contractile activity² or as an enhanced damage response in people with Duchenne muscular dystrophy who exercise³.

The isolated, intact mature muscle-fibre studies by Menke and Jockusch were virtually free of extracellular matrix as a result of the collagenase treatment used in the isolation procedure. Surrounding muscle cells and pericellular fibrous tissue may well exert a restraining effect on muscle cells *in vivo*, preventing the substantial distortions of the membrane and hypercontraction apparent in hypo-osmotic shock to isolated muscle fibres. Further, the nature of the mechanism of cell death is unknown in hypo-osmotic shock, but the stress on the membrane presumably involves a substantial increase in internal pressure in the myofibres as extrusions (blebs) on the cell membrane are formed, a type of stress not presented to the muscle fibre in the models of contractile activity induced by damage.

Interestingly, the progression of the damage to the isolated fibres described by Menke and Jockusch is not unique to hypo-osmotic shock, but is qualitatively similar to that which occurs in isolated fibres under various forms of stress, including trauma and deliberate elevation of intracellular calcium with the calcium ionophore⁴. It therefore seems necessary to determine whether the decreased stability of mdx muscle found by Menke and Jockusch is a consequence of the hypo-osmotic shock or of the nature of the model system used. Although the results of Menke and Jockusch's report are interesting and potentially important, the 'fragile membrane' theory is by no means proven; for an abnormality such as increased fragility to be important in the pathogenesis of Duchenne and Becker muscular dystrophy, it must be apparent under physiological conditions.

M. J. JACKSON

A. McARDLE

R. H. T. EDWARDS

Muscle Research Centre,
Department of Medicine,
University of Liverpool,
PO Box 147,
Liverpool L69 3BX, UK

D. A. JONES

Department of Medicine,
University College,
London WC1E 6BT, UK

1. Menke A. & Jockusch H. *Nature* **349**, 69–71 (1991).

2. Jackson M. J., McArdle A. & Edwards R. H. T. *J. Neurol Sci.* **98** (suppl.), 239 (1990).

3. Jackson M. J. et al. *Muscle & Nerve* **10**, 15–21 (1987).

4. Zuurveld J. G. E. M., Veerkamp J. H. and Wirtz P. *Muscle & Nerve* **8**, 750–759 (1985).

1. Kraetschmer, W., Lamb, L. D., Fostiropoulos, K. & Huffman, D. *Nature* **347**, 354–358 (1990).

2. Kelly, B. T. *Physics of Graphite* **74**, (Applied Science Publishers, Englewood, New Jersey, 1981).

3. Barrett, C. S. & Massalski, T. B. *Structure of Metals* **627**, (McGraw-Hill Book Co., New York, 1966).

4. McSkimin, H. J. & Bond, W. L. *Phys. Rev.* **105**, 116–121 (1957).

5. Westbrook, J. H. & Conrad, H. *The Science of Hardness Testing and its Research Applications* (American Society for Metals, Metals Park, Ohio, 1973).

6. Liu, A. Y. & Cohen, M. L. *Science* **245**, 841–842 (1989).

They also serve

William H. Press

The Joy of Insight: Passions of a Physicist. By Victor Weisskopf. *Basic Books*: 1991. Pp.336. \$24.95.

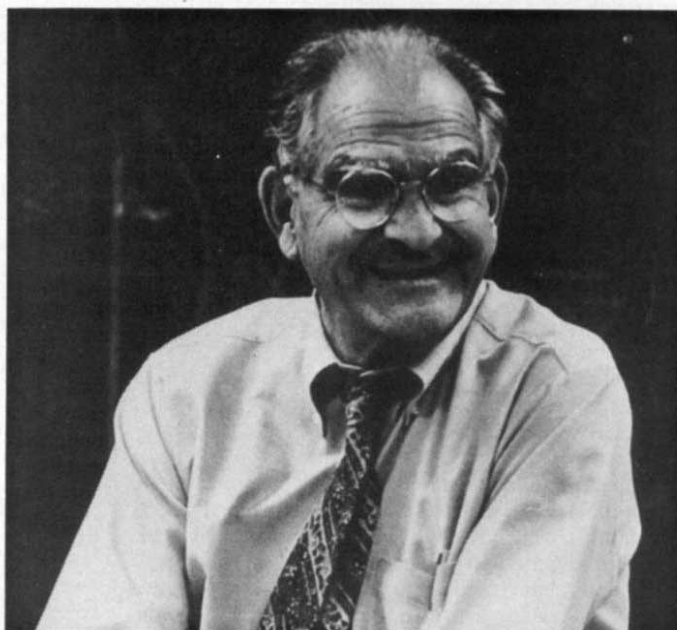
VICTOR Weisskopf, is a fortunate and happy individual, and he does not mind telling us so. Born in Vienna in 1908 to a well-to-do and intensely cultured Jewish family, Weisskopf lived through the Nazi destruction of European culture and science without harm to himself or any of his immediate family. An emigré in Rochester, then a Manhattan Project group leader at Los Alamos, and later an active citizen in the United States on issues of arms control and nuclear disarmament, he returned to Europe to play a significant role in the re-emergence of large-scale European science at CERN in the 1960s. Subsequently he has been a much-honoured senior statesman of science, a fixture of the Cambridge, Massachusetts, and world scenes, a man who summers in his '*petit paradis*', a chalet on the slope of the Jura Mountains with a view of Mont Blanc.

And what physicist would not trade 10,000 citations in the *Science Citation Index* for a chance to have been where Weisskopf was, so often at the very centre of the action of twentieth-century physics? As a student in gymnasium, he read voraciously the popular accounts of the then infant quantum mechanics, wrote a letter to Planck and received a postcard reply (a wrong explanation, we now know). Hans Thirring, his undergraduate mentor in Vienna, sent him on to Göttingen, where he studied under Max Born, James Franck, Paul Ehrenfest and Richard Courant, among others. For recreation there were trips to Berlin to see the musical theatre of Brecht and Weill and their followers, or a week in Jena listening to Furtwängler conducting Brahms.

In 1931, doctorate in hand, Weisskopf went to Leipzig, to work with Heisenberg (alongside Yoshio Nishina and Felix Bloch); Schrödinger invited him to Berlin as his assistant. "Often I saw the Nazi gangs . . . beat up Jewish students or those who looked Jewish . . . More than once I had to pull one of the boys under attack into my office so that he could escape through the back door." Weisskopf spent a year in Copenhagen with Niels Bohr (where he met his future wife, Ellen), then several years as Pauli's assistant in Zürich. Expelled by the Swiss *Fremdenpolizei* for alleged leftist connections (a bureaucratic fabrication, he says), Viki and

Ellen spent some months in Copenhagen and Vienna, and crossed the Atlantic in 1937.

After the war and Los Alamos, Weisskopf was recruited to the Massachusetts Institute of Technology, where he became a powerful department head, wrote a classic textbook on theoretical nuclear physics, and supervised graduate students and postdoctoral fellows including Sid Drell, Murray Gell-



Victor Weisskopf — a life fulfilled.

Mann, David Jackson and Kurt Gottfried. Weisskopf was director-general of CERN from 1961 to 1965, during which time the decision to build a colliding-beam accelerator was taken.

When we read the autobiographical story of a charmed life, we are not surprised to find it told by a charmer. Weisskopf is the very model of the modern, liberal, politically conscious and administratively adept, twentieth-century physicist, a major-domo of 'big science'. He did about everything there was to do, except make a major scientific discovery himself. Weisskopf's role has always been that of the facilitator, the brilliant and fair-minded administrator, the writer of textbooks and review articles and the inspiring lecturer capable of explaining the big picture. "It is regrettable," he writes, "that among scientists the presentation of ideas is not as highly valued as the creation of ideas. This is in stark contrast to music, where the performer is a partner equal to the composer."

If it did not ring so true, Weisskopf's

memoir, *The Joy of Insight*, would seem intolerably smug. For example, "We were an unusual community, an international crowd of extremely creative people, and even our informal social gatherings were extraordinarily stimulating and interesting." (This could refer to any period in his life; it happens to be at Los Alamos.) When a rare cloud does shade this author's sunny disposition, it often has something to do with the question, what are sources of self-esteem for the almost-great scientist who has missed forever the young person's chance of making a truly important discovery? How many honours from professional societies and governments, how many good friends, splendid concert or theatre evenings, would one trade for a really important 'joy of insight' — important enough to make one's name immortal?

Weisskopf recognizes that, before the war, he might have been able to make discoveries in quantum field theory that were in the event made by others after it. If only he had been a bit more mathematical. If only he had worked a bit harder, or loved music and socializing a bit less. But, on balance, he has few regrets.

In one of this book's rare sardonic passages, Weisskopf writes: "There are three periods in the life of a successful theoretical physicist. The first, when one is young, is the time of hard work on new ideas . . . In the second period one has become reasonably well known and is invited to give talks at various universities and at conferences. One still contributes valuable original work, but more often with the help of younger colleagues or by suggesting a line of

research to them. In the third period one receives invitations to give general talks surveying the whole field through the light of experience. A little later one is invited to deliver memorial speeches about deceased colleagues or talks on the philosophical and political aspects of science. Finally, all one is asked to do is give after-dinner speeches."

Such shadows aside, Vienna, Göttingen, Copenhagen, Los Alamos, Cambridge and Geneva all make for a warm and engaging collection of stories, especially when they have been polished to a high sheen by years of retelling, and are now finally written down from the perspective of a mellow and much honoured old age. In a final chapter, the author discusses his personal, ranked choices of the greatest musical compositions and, especially, operas. □

William H. Press is in the Center of Astrophysics, Harvard University, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

Molecular oncology

Antony J. Michalski

Genes and Cancer. Edited by Desmond Carney and Karol Sikora. Wiley: 1990. Pp. 348. £24.95, \$49.50.

Molecular Genetics and Cancer Diagnosis. By Jeffrey Crossman. Elsevier: 1990. Pp. 465. Dfl165, \$69.50.

THE last two decades have seen an exponential rise in the amount and diversity of research in molecular biology. More recently, results of this technology have made an impact on clinical practice. Recombinant DNA techniques are used to produce growth hormone, insulin and haemopoietic growth factors. Identification of relevant restriction-fragment length polymorphisms is now routine in carrier detection and prenatal diagnosis of genetic diseases. The polymerase chain reaction is being used to amplify virus-specific sequences in the diagnosis of various virus infections, for example, human immunodeficiency virus in neonates. The genetic events responsible for tumorigenesis in retinoblastoma and colon cancer have now been identified and huge steps are being made in understanding the molecular pathogenesis of Wilms' tumour, Burkitts lymphoma and many other malignancies. No longer can clinicians and 'nonmolecular' scientists remain ignorant of the basic principles, techniques and jargon of molecular biology and they need a succinct and understandable introduction to this highly complex field. *Genes and Cancer* edited by Desmond Carney and Karol Sikora, and *Molecular Genetics and Cancer Diagnosis* by Jeffrey Crossman are aimed at a general readership and attempt to fulfil this role, though in different ways.

Genes and Cancer is divided into six sections covering oncogenes, gene control, tumour biology, growth factors and receptors, drug resistance and the use of antibodies for cancer therapy. The sections contain between four and eight articles of around ten pages in length and are followed by short but well-chosen lists of further reading. The authorship is largely UK based but there are contributions from the United States and Japan. The strength of the book is its broad scope and the many chapters that provide excellent reviews of a topic with the minimum of extraneous detail. In any evolving field, deciding on the balance between communicating hard data and speculative interpretation is difficult, but some chapters are made less digestible for a general readership by the inclusion of the full sequence of an oncogene or the configuration of its protein product. It is impossible for a book of this size to be comprehensive but I was surprised not to find a single chapter on 'tumour-suppressor' genes. Similarly, although

paediatric malignancy numerically accounts for a tiny proportion of all cancer, it could have received more space if only to use retinoblastoma as a vehicle to discuss tumour-suppressor genes or Wilms' tumour to discuss genomic imprinting.

By contrast *Molecular Genetics in Cancer Diagnosis* is narrower in scope but more cohesive, with themes progressively developed by a variety of authors. It is divided into 25 chapters with up to 12 subheadings in each. The authorship is virtually all North American and has chosen to give comprehensive, but rather daunting, reading lists at the end of each chapter. The book concentrates much more on the application of formal molecular biology to oncology and does not attempt to cover current or future treatment modalities in any detail. The first eight chapters form a well-written, succinct introduction to the genetic basis of cancer, molecular techniques, oncogenes, tumour-suppressor genes, growth factors, metastasis and drug

resistance. The chapter on the cytogenetic basis for molecular analysis of malignancies is particularly useful. The remaining chapters discuss the molecular pathology of selected malignancies. The majority are 'disease based' explaining the relevance of cytogenetic abnormalities, oncogene activation and loss of tumour-suppressor gene function to the tumour under discussion. There are several chapters on haematological malignancies but the solid tumours receive less complete coverage in terms of the variety of diagnoses considered, though the articles themselves are amongst the best in the book. Although some of the chapters are more detailed than others, and a few communicate data in a more philatelic than interpretative way, this is a readable and informative book. □

Antony J. Michalski is in the Department of Haematology and Oncology, Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK.

Awaiting tests

Neil Turok

Particle Physics and Inflationary Cosmology. By Andrei D. Linde. Harwood: 1990. Pp.362. £32, \$60.

THE idea that one or more phase transitions occurred in the early Universe lies behind most of the theoretical work in cosmology over the last two decades. *Particle Physics and Inflationary Cosmology* is one of the first books devoted to a detailed technical treatment. Its author has been intimately involved in the subject since its inception, and has consistently been one of the most active and imaginative workers in the field. In the early 1970s, Andrei Linde was one of the first to realize that the standard electroweak theory predicts such a phase transition, and in the early 1980s played a leading role developing the inflationary model, through which the dynamics of a grand unified phase transition could resolve many of the basic conundrums of cosmology.

Linde has since then pursued new models of inflation, and their consequences, with great vigour. His particular strength is in finding simple physical arguments to clarify and explain the very complex processes involved. This comes through strongly in his book, which provides simple but quite deep treatments of many of the more technical issues — effective potentials, spontaneous symmetry breaking, the generation of fluctuations. It should serve as both a textbook and a standard reference on inflation for years to come. It is packed with original and intuitive viewpoints on finite-temperature field theory, inflation and quantum gravity. The emphasis is always on simple physical pictures and rough scaling arguments rather than on rigorous mathematical formalism. This can be a little frustrating — for example in the

discussion of quantum gravitational effects, for which there is as yet no consistent theory, but which are treated in a crude intuitive manner. But Linde's point of view is always simple and stimulating.

The first half of the book is a nice review of phase transitions in field theories, including those in the standard model, bubble nucleation and topological defects. The main emphasis in the second half of the book is on an enthusiastic account of the many versions of inflation and in particular of two models: 'new inflation' and 'chaotic inflation'. The latter has come to be seen as probably the simplest embodiment of the inflationary hypothesis. But chaotic inflation is still on rather shaky footing: the initial conditions 'before' it are discussed using very rough arguments close to the Planck era, where unknown quantum gravity effects are undoubtedly important. This vagueness lends a note of unreality to the discussion. The book closes with a long heuristic discussion of quantum cosmology, the 'self-regenerating universe' which is a development of the chaotic model, and the anthropic principle. Somewhat disappointingly it thus ends firmly in the realm of philosophy rather than experiment.

My main criticism of the book is that it contains barely a mention of the observational data. Linde's primary concern is with the basic aesthetic problems of the hot Big Bang, and their resolution in theories of inflation, for which he is a very strong and able proponent. Inflation provides a simple dynamical mechanism through which a small lumpy 'typical' universe could become large, flat and smooth like our own. At present there is no other mechanism known which does this, barring the supposition of special initial conditions, which may in the end always resolve these issues. Furthermore, inflation provides a mechanism for the generation of small-amplitude primordial 'wrinkles' on the universe, which later grow

under gravitational instability to form stars, galaxies and superclusters.

These successes have led some to speak of the 'inflationary paradigm', and to presume that the basic ideas of inflation are correct. In my view at least, that is a little premature. Experiment is surely an integral part of scientific progress, and our theories are worth very little if they are not testable. Of course there are fundamental limits to this in cosmology due simply to our inability to repeat or observe the early stages of the Big Bang. It is also conceivable that a theory might be so mathematically compelling that we might accept its validity even in situations far from those experimentally accessible. Such is the case, for example, in general relativity where most practitioners believe the theory correctly describes what would happen if one fell across the Schwarzschild horizon of a black hole. The various models of inflation are very far from this level of aestheticism, a fact their great number testifies to. This does not mean they are wrong — the standard model of particle physics, which is in wonderful agreement with a huge amount of real data, is actually pretty ugly. But models of inflation do require experimental verification.

The exciting thing about cosmology today, and what is really new about it as a science, is that this is no longer out of the question. The predictions of inflation for the form of the density fluctuations in the universe are eminently testable. First, by observations of the galaxy distribution on very large scales. Recent observations have actually shown significant disagreement with the inflationary predictions. Second, and probably more 'cleanly', inflation makes clear predictions for the precise form of the microwave-background anisotropy. Microwave-background experiments have reached a level of precision where they are close to ruling out or confirming most cosmological models for the large-scale structure of the universe, including those discussed in Linde's book.

Despite its neglect of these issues, I warmly recommend *Particle Physics and Inflationary Cosmology* both to serious students and researchers in the field. Linde's bubbly enthusiasm for a hundred theories and models at once provides both a stimulation for theorists and a challenge for experimentalists. □

Neil Turok is in the Department of Physics, Princeton University, Jadwin Hall, PO Box 708, Princeton, New Jersey 08544, USA.

New in paperback

■ *Bivalve Filter Feeding: Hydrodynamics, Bioenergetics, Physiology and Ecology* by C. Barker Jorgensen focuses on filter-feeding bivalves and the ecological and physiological factors that determine the rate of filtration. Published by Olsen and Olsen at £22.

■ From Oxford University Press comes *Quantum Description of High Resolution NMR in Liquids* by Maurice Goldman for those with no working knowledge of quantum calculations. Price is £17.50, \$35.

Central player

Richard I. Gumport

Nucleic Acids in Chemistry and Biology.

Edited by G. Michael Blackburn and Michael J. Gait. Oxford University Press: 1990. Pp.446. Hbk £50, \$115; Pbk £22.50, \$49.95.

A COROLLARY to the oft-stated theorem that 'the future is biology' is that 'the future of biology is chemistry'. For students wishing to begin acquiring the knowledge necessary to fulfil the corollary, *Nucleic Acids in Chemistry and Biology* is an excellent textbook. Although the text will be most useful to advanced undergraduate and beginning graduate students, practising chemists and molecular biologists who wish to garner a molecular understanding of the nucleic acids will also benefit from reading it. Students will require preparation in organic and physical chemistry and biochemistry to exploit the book fully, but little previous exposure to molecular biology is required. By stressing the pedagogical virtues of the text in this review, I do not mean to indicate that scientists actively engaged in research on nucleic acids will not learn by perusing the text, for it collects and presents information according to the judgements and perspectives of authors who are major contributors in their fields.

After a brief introduction in which the history of nucleic-acids research is interestingly recounted, the editors provide nine chapters written by themselves and selected specialists from Britain and the United States. The writing is clear and concise, and the editors have generally imposed a uniformity of style and depth of coverage upon the contributors that leads to overall coherence. Key words are in boldface and a relatively complete glossary with succinct definitions is provided. The major sections of each chapter have boxed, concise summaries that highlight the essential content. The many chemical structures and drawn-out reactions exemplify the emphasis of the text. Numerous well-conceived figures, some coloured in two tones for clarity and others depicted for stereoscopic viewing, contribute to the presentation. A comprehensive index further indicates the authors' intent to facilitate the use of the book by students. References to authoritative reviews and original research literature for each section of the chapters are provided. In many instances it is unfortunate that these are not designated in the text when the referenced topic is covered.

The body of the text begins with a consideration of DNA and RNA structure; the material proceeds from the heterocyclic bases through chromatin. The families of DNA duplexes, 'unusual' DNA structures, that is, cruciforms, curved DNA, supercoiling, and triple-strand nucleic-acid structures as well as a brief overview of the dynamics of nucleic-acid structures are covered. In the

longest and most detailed chapter, the chemical syntheses of nucleosides, nucleotides and oligoribonucleotides and oligodeoxyribonucleotides are presented. The use of enzymes in synthesizing and joining oligonucleotides is also outlined. A short chapter that provides an overview of the biosynthesis of nucleotides and nucleic acids ends with a consideration of how drugs are used to inhibit some of these reactions. The following two chapters on DNA and RNA sequence information and transmission are presented at a very basic and introductory level. Gene structure, gene expression and its regulation, RNA synthesis and catalysis, and protein synthesis are outlined. The chapters will be most useful to those with chemical backgrounds who wish to begin to understand why there is such emphasis on the nucleic acids in modern biological research.

The remainder of the book is concerned with specific aspects of nucleic-acid chemistry. A relatively detailed chapter considers the chemistry of covalent interactions between small molecules and nucleic acids. A brief overview of the biological consequences of such modifications and of DNA repair concludes this topic. The chapter considering the noncovalent interactions of small molecules with nucleic acids covers groove-binders, intercalators, bisintercalators and DNA cleavage reagents. A chapter devoted to protein-DNA interactions starts with the forces that stabilize the complexes and proceeds through nonspecific to sequence-specific interactions. Structures of DNA-binding proteins, repressor-operator complexes, and an enzyme-DNA complex derived from X-ray diffraction studies are presented to explain the various mechanisms by which specificity arises. The contributions of using base modifications to study specific interactions are also presented. The final chapter is a basic introduction to genetic recombination with the emphasis on DNA and not the battery of enzymes and proteins that effect the processes. Homologous, site-specific, illegitimate and transposition recombination are presented. The chapter ends with an outline of cloning, recombinant DNA technology, *in vitro* mutagenesis, and a consideration of gene replacement.

There are many current textbooks available on molecular biology, but few that focus on the chemistry of a central player — the nucleic acids. This text fills that void with introductory, authoritative and up-to-date coverage. In spite of a few niggling errors that are inevitable in first printings and the occasional overstatements made during the necessary generalizations of an introductory work, this text will be extremely useful. Many students may be started along exciting and fruitful paths and some of their professors have their research itineraries modified as a result of studying this text. □

Richard I. Gumport is in the Department of Biochemistry, University of Illinois, Urbana, Illinois 61801, USA.

Falling victim to politics

Stuart L. Pimm

The Expendable Future: US Politics and the Protection of Biological Diversity.

By Richard Tobin. *Duke University Press*: 1991. Pp.325. Hbk £42.50, \$45; pbk £17, \$18.75

As species extinction accelerates worldwide, we look in despair at developing, tropical, countries that sacrifice biological riches for short-term economic gains, or even poorer reasons. Surely, in a country like the United States, with an affluent and well-educated population, circumstances should be different. As Richard Tobin explains at the beginning of *The Expendable Future*, the great majority of Americans believe that we should prevent extinctions. He quickly notices that surveys indicate that about 60 per cent of people with these views, also feel that "modifying the environment for human use seldom causes serious problems", and that our species "was created to rule the rest of nature". Thus, there is likely to be a wide gap between expectations and reality. Tobin documents that gap with clear writing and carefully constructed expectations of what should not, but inevitable do, fall victim to politics.

The first expectation, given the majority view, is that the US agency mainly responsible for protecting diversity, the Fish and Wildlife Service (FWS), should be well-placed within the administrative structure. In fact, it is buried within a bureaucracy, surrounded by agencies largely devoted to the subjugation of nature. Whatever its location, the FWS can assert itself in several ways. The most crucial stage in the protection of endangered species is the first — the listing of species on the brink of extinction. This listing process began in the early 1970s with high expectations. In 1975, with the help of the Smithsonian Institution, over 3,000 species of plants were suggested as possible candidates. The Endangered Species Act, however, contained a short, apparently innocuous paragraph that required all federally-funded projects to not jeopardize endangered species. FWS listed the snail darter, (a small fish), challenged the Tennessee Valley Authority's construction of the already highly controversial Tellico Dam, and won in the US Supreme Court. The local senator, Howard Baker, had to introduce special legislation to permit the dam to be finished. Listing species became a difficult and very slow process.

The FWS maintains that its decisions to list species are based on "the best scientific and commercial data available" — just as the law requires. Listing should be on the basis of threat, not on political expediency, or whether the animal or plant is small and

slimy or cuddly and large. Globally endangered species (particularly if in a monotypic genus), should receive priority over locally endangered populations of an otherwise flourishing species. Listing should be more frequent in Hawaii where most of the endangered species are to be found, than in the continental United States. In practice, the species listed tend to be local populations of large birds or mammals, in mainland areas where they cause few impediments.

In addition to listing, the FWS can designate critical habitats for species. Again expectations were high, and again, the resulting controversy often considerable, and the subsequent actions few. Rather more tangible is

should we ecologists get from this book? First, the book infuses a sense of political reality: the federal government will not save small ugly invertebrates with funny sounding names; we should seek help for these species elsewhere. Second, in the decisions to save species and the habitats that house them, ecologists could do much more to become informed about where endangered species are to be found. Spending over \$20 million to save the Mississippi sandhill crane (a subspecies) and more on bald eagles than on any other single species, while very much less to save the very large number of critically endangered species of plants in the tropical forests of US states and territories (Hawaii,



Symbol of America — the bald eagle is in no immediate danger yet receives more financial assistance than any other species. Picture from *Hawks, Eagles and Falcons of North America* by Paul A. Johnsgard (Smithsonian Institution Press, \$45, £29.95).

the acquisition of land to house endangered species. Tobin states that "considerable amounts of money have been devoted to lease purchase habitats for subspecies facing a low or medium threat and having doubtful prospects for recovery." Bald eagles have done very well financially, despite a FWS assessment that the species is in no immediate danger. Finally, there are the recovery plans — the actions proscribed to return species to better times. Again, successes are few.

The author is a political scientist and presents the organizational history, political decisions, and changing laws and administrative structures in ways that as an ecologist I found compelling and interesting. He concludes with recommendations for the politicians, lawyers and administrators. But what

Guam and other Pacific areas, Puerto Rico and other Caribbean areas) is inexcusably provincial. How can we save tropical forests worldwide if we cannot save them at home? Ecologists need to know where the diversity is, where it is threatened, and be more assertive in their priorities. Finally, there is considerable scientific uncertainty about how to determine the relative threats to species and the likelihood of recovery. Without better ecology, endangered species are going to remain potential victims of political whims. □

Stuart L. Pimm is at the Department of Zoology and Graduate Program in Ecology, The University of Tennessee, Knoxville, Tennessee, 37996, USA.

Food web patterns and their consequences

Stuart L. Pimm, John H. Lawton & Joel E. Cohen

A food web is a map that describes which kinds of organisms in a community eat which other kinds. A web helps picture how a community is put together and how it works. Although webs were often initially reported in despair at ever understanding ecological complexity, recently discovered widespread patterns in the shapes of webs, and theoretical explanations for these patterns, indicate that webs are orderly and intelligible, and have some foreseeable consequences for the dynamics of communities.

FOOD webs are the road-maps through Darwin's famous 'entangled bank'¹ and have a long history in ecology². Like maps of unfamiliar ground, food webs appear bewilderingly complex. They were often published to make just that point. Yet recent studies have shown that food webs from a wide range of terrestrial, freshwater, and marine communities share a remarkable list of patterns. Current research concentrates on how many independent patterns there are, how they may be described quantitatively, why the patterns are so general, and what are the consequences of these patterns for the dynamics of a community and its constituent species. Just as any map omits details, most published webs omit predation on minor species, the quantities of food consumed, the chemical composition and temporal variation of the flows and many other details. Published webs are also of very variable quality. These omissions and problems are causes of concern, but on present evidence do not present insurmountable difficulties.

The data

Food web complexity is exemplified in Fig. 1. (This web³ pre-dates Shelford's study⁴, which is often credited as the first published web, by one year.) The community embracing the boll weevil is certainly complex, but not hopelessly so; indeed it raises several issues of contemporary interest. For instance, what level of taxonomic resolution is appropriate for work of this kind (most taxa in Fig. 1 are highly aggregated, but some are identified to species); and how are the boundaries of the studied system to be delineated?

Three kinds of food webs are published. A source web includes one or more kinds of organisms, the organisms that eat them, their predators and so on (for example, the part of Fig. 1 based on the cotton plant). A sink web describes one or more kinds of organisms, the organisms they eat, plus their prey, and so on. A community web is defined by picking, within a habitat or set of habitats, a group of species without regard to the eating relations between them, and working out who eats whom⁵. Reference 6 compiles 113 community webs and most of these webs contain between 5 and 50 species. We focus here on

community webs⁵⁻⁷, an example of which is shown in Fig. 2. Terms used to describe web features and the information that can be extracted from them are defined in Boxes 1 and 2. Over 200 community webs have been reported in the literature⁷ (available in machine-readable form through ECOWEB⁷) embracing a wide variety of habitats, locations and taxa. The analyses of these webs are usually in terms of trophic species—a set of organisms with identical prey species and identical predators, within the level of resolution used in the study. Figure 1 is unusual in that it indicates the number of species grouped into a trophic species; there are, for example, 12 species grouped under the trophic species of 'leaf worm parasite'. In general, a trophic species may correspond to a set of biological species, to a single biological species or to a single life-stage of a species. We shall use 'species' to mean 'trophic species' throughout.

Features common to all webs

Box 1 explains the statistics commonly derived from a web and also gives the particular statistics derived from the *Nepenthes* web. Two associated graphs, the predator overlap graph (also called the trophic niche overlap graph⁵ or, by graph theorists, the competition graph) and the prey overlap graph⁸ are derived from an analysis of the web (Box 2 and Figs 3 and 4). Several features are common to the set of webs studied so far, a set that has widely varying numbers of species and contains webs from many different habitats. Common features (see Box 3) include the average proportions of top, intermediate and basal species, the ratio of predatory species to species of prey, the proportions of trophic linkages between different kinds of species, the length of food chains, the absence of compartments within a habitat, and complex patterns in the topological relationships between predators and prey elucidated by the predator and prey overlap graphs.

Are the patterns real?

Are these general patterns artefacts? There are good reasons for concern about the quality of data in published webs. Communities often contain thousands of species. Because published

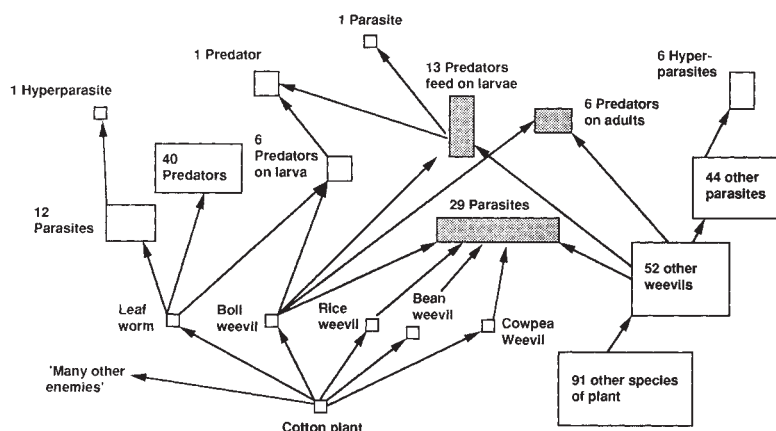


FIG. 1 The food web associated with the cotton plant and the boll weevil (after ref. 3). The cotton is attacked by 'many other enemies' than those specified. Moreover, the predators shown in the three shaded boxes are known to attack species other than those shown in the figure.

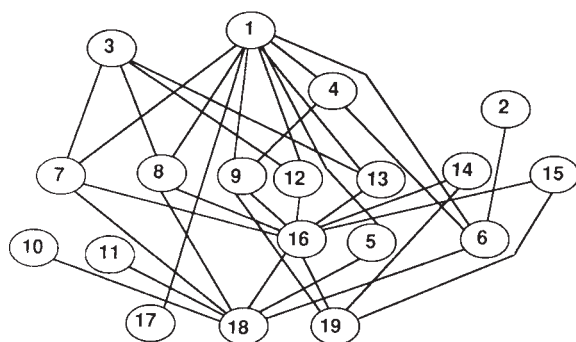


FIG. 2 A food web of the insects in the pitcher plant *Nepenthes albomarginata* in West Malaysia (from ref. 16). Each line represents a trophic linkage; predators are higher in the figure than their prey. Key: (1) *Misumenops nepenthicola*; (2) Encyrtid (near *Trachinaephagus*); (3) *Toxorhynchites klossi*; (4) *Lestodiplosis syringopais*; (5) *Megaselia* sp. (?*nepenthina*); (6) *Endonepenthia schuitemakeri*; (7) *Triperoides tenax*; (8) *T. bambusa*; (9) *Dasyhelea nepenthicola*; (10) *Nepenthosyrphus* sp.; (11) *Pierretia urceola*; (12) *Culex curtisipalis*; (13) *C. lucaris*; (14) *Anotidae* sp. 1; (15) *Anotidae* sp. 2; (16) bacteria and protozoa; (17) live insects; (18) recently drowned insects; (19) older organic debris.

webs include only tens of trophic species (>100 is exceptional), they are either highly aggregated or represent only a tiny part of the entire system. Aggregation is rife in many published webs; moreover, aggregation varies in extent from web to web, and at different positions in the same web⁹. Even when webs are detailed enough for most of their elements to be single biological species (for example Fig. 2), the linkages are less often based on experimental evidence than on casual observations; only

those linkages that are 'artistically convenient'¹⁰ may appear in the figure. In recent studies we detect a counter-tendency to report every linkage, no matter how minor. We appeal for studies where the frequencies of feeding encounters are reported, so that objective criteria may be used to include or exclude linkages.

There are a number of rejoinders to the criticisms that the data are too aggregated and unreliable to reveal meaningful patterns. Possible artefacts ought not to apply equally to webs from special microhabitats described in detail and to those more aggregated webs summarizing interactions throughout larger communities. Present evidence suggests that most patterns appear across all webs. Moreover, artefacts should be more obvious in the highly aggregated webs. Indeed, the ratio of species of predators to species of prey is sensitive to the greater aggregation of species at lower trophic levels⁹, and the ratio itself has dropped as better data have become available^{11,12}. Nevertheless, the average proportions of species at different levels change only slowly or not at all across webs differing widely in numbers of species and degree of aggregation, though the variation in these proportions is large. Detailed webs can be progressively aggregated. For instance in Fig. 2, the three pairs of biological species (10 and 11; 12 and 13; 14 and 15) that have the same sets of predators and prey could be joined into three trophic species. The web could be further aggregated by joining 7, 8, 12 and 13, which share similar species of predators and prey. Aggregation of webs using trophic criteria affects webs' properties only slightly¹³; in contrast, progressive aggregation by taxonomic affinities alters webs' properties more rapidly (K. Schoenly and G. Sugihara, manuscript in preparation).

Recent studies to address limitations in the data (for example, refs 14–18) have generally confirmed the patterns. The pattern

BOX 1. Definitions of web features and analysis of the *Nepenthes* web

Top predators are species on which nothing else in the web feeds, basal species feed on nothing within the web, and intermediate species have both predators and prey within the web. Linkages are trophic connections between species.

Cycles occur where, for example *j* eats *i* and *i* eats *j*, or *i* eats *j*, *j* eats *k* and *k* eats *i*. Cannibalism is a cycle where one species feeds upon itself.

Connectance is the number of realized trophic interactions divided by the number of possible interactions. Within this definition are several closely related variants. The number of possible interactions may be S^2 if cannibalistic interactions are counted, $S(S-1)$ if only inter-specific interactions are counted. We may assume that for each linkage between predator and prey there are two interactions: the effect of the predator on the prey, and the effect of the prey on the predator. Alternatively, we may assume that either of these two effects is dynamically zero. Typically, connectance is calculated from twice the number of observed linkages divided by $S(S-1)$ for the number of possible interactions. Hence, when the number of links per species (linkage density, d), is constant, connectance declines approximately hyperbolically with increasing number of species.

The trophic level of a species in webs of linear structure is one more than the chain length, that is, the number of linkages between it and the basal species in the web (which may be plants or detritus). In reality, species may trace linkages to basal species along food chain pathways of different lengths.

Omnivores are species that feed on more than one trophic level⁹. Omnivores blur the distinction between trophic levels, because top predators may trace linkages to basal species through food chains of different lengths. Omnivory can involve either different food chains or the same food chain. In the latter case, for example, an omnivore may feed on both a prey species and that prey species' prey species. Other definitions of omnivory are possible²⁰.

Compartments exist when linkages are few (or weak) between groups of species and common (or strong) within those groups.

Species analysis

Number of top predators is 7: species **1, 2, 3, 10, 11, 14, 15**.

Number of basal species is 3: species **17, 18, 19**.

Number of intermediate species is 9: the rest.

Number of linkages: intermediate to top 14; basal to top 4; intermediate to intermediate 8; basal to intermediate 7.

Cycles. There are no cycles in this web.

Connectance and linkage density.

Connectance = 0.19 (calculated as twice the number of linkages divided by $S(S-1)$). Linkage density = $33/19 = 1.74$.

Chain lengths

Minimum 1: for example, species **1** to **17, 15** to **19**. Maximum 4: species **1** to **4** to **9** to **16** to **19**.

Modal number for each top predator: species (Mode)

1 (3); **2** (2); **3** (3); **10** (1); **11** (1); **14** (2); **15** (2).

Omnivory. In this web omnivory occurs via different chains, for example, species **1** to **17** and **1** to **18** and **19** via longer pathways, and also within chains, for example, species **1** to **5** and **1** to **4** to **5** or species **9** to **19** and **9** to **16** to **19**.

Compartments

None in this web.

There may be a correlation between the number of species of prey each intermediate species exploits and the number of species of predators that exploit the intermediate species

Intermediate species	4	5	6	7	8	9	12	13	16
No. of prey	2	1	1	2	2	2	1	1	2
No. of predators	1	1	3	2	2	2	2	2	7

least likely to survive detailed scrutiny involves the linkage density, d , which is the number of trophic linkages L per (trophic) species S . Averaged over webs in the range from 3 to 48 species, the average number of linkages ($E(L)$) is roughly twice the number of species⁶ in any given web (that is, $E(L) = 2S$; $d = 2$). This is equivalent to a hyperbolic relation between connectivity C and species richness S (see linkage density in Box 3). In the original description of this pattern⁶ it was noted that a power-law $E(L) = kS^{1+\epsilon}$, for some small positive ϵ , was also a viable description of the data, and that future data on webs with large numbers of species would have to distinguish the alternatives. With a few larger webs in hand, a power law with ϵ probably between 0.3 and 0.4 indeed seems reasonable^{6,19}.

What the patterns do not show

Although most analyses ignore the problem, web structure varies over time and space^{15,20}. A useful distinction²¹ is between cumulative webs gathered over many occasions (the majority of the published webs) and time-specific webs. In most of 16 habitats with reported time-specific webs, the percentages of top, intermediate and basal species fluctuated widely, generating appreciable variation in predator-to-prey ratios, mean chain length, and linkage density²¹. Within any one habitat, cumulative webs usually overestimate linkage density and underestimate the percentage of basal species relative to time-specific snapshots²¹. Web statistics are sensitive to occasional rare species (typically large, generalized predators) encountered on only some sampling occasions.

Comparable systems involving related taxa and/or similar habitats reveal differences in web statistics from place to place within one region¹⁵ and locally within a single habitat¹⁴. The significance of these differences is not well understood.

The question of how to define the boundaries of a web is particularly vexing. Most investigators tacitly admit that members of most webs, particularly those higher in the food chain, feed outside the system studied. Published webs are reticulate, with no evidence of compartments (Box 3) except, possibly, fuzzy ones at habitat boundaries⁹. Given the apparently arbitrary nature of the boundaries of most studies, one conclusion must be that web statistics are roughly independent of spatial scale. This leaves open the question of whether spatial and temporal variation in web structure are interchangeable²¹. In certain physical systems, a single system observed over a long time period reveals variation that is identical to that observed in many systems of the same kind simultaneously but at different places. Interchangeability of temporal and spatial variability is called ergodicity; we do not know whether web statistics are ergodic.

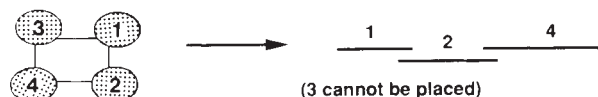
Differences between systems

Although webs have common features across a wide range of habitats and taxa, there is some variation. Counter to earlier predictions, average food-chain lengths do not seem to differ greatly among ecosystems with very different primary productivities^{6,9}. There are only slight differences in chain lengths between communities where the consumers are often vertebrates compared with those where the consumers are all invertebrates^{22,23}. Chains may be shorter in small habitat patches. Thus, on very small islands predators present on larger islands may be absent, and small ponds may lack the predators of larger lakes¹⁹. In some special microhabitats, chains are shorter in areas with frequent natural or experimental disturbances²⁴. Reported chains are shorter in two-dimensional habitats such as grasslands and the intertidal than in three-dimensional habitats such as forests or the water column of lakes or oceans⁶. Intertidal ecologists, however, do not record the predatory fish that visit their systems at high tide, whereas those working below the low tide mark do. Hence, different average chain lengths may be an artefact of different academic traditions or, as discussed in ref. 19, a genuine difference in the effects fish have on bottom-living versus free-swimming prey. We caution against

BOX 2. Predator and prey overlap graphs

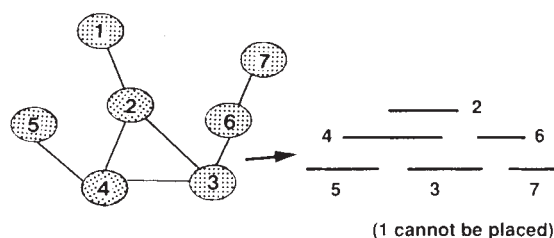
From a given web, we can obtain another graph, called the predator overlap graph. In this graph, an edge connects predatory species that share one or more species of prey in their diets. There may be more than one set of connected predators: these sets are called components.

Predator overlap graphs are rigid circuit, when every circuit around four or more predators is divided into triangles by one or more edges across the circuit. Below is a hypothetical overlap graph that is not rigid circuit, it would be if an edge joined predators 1 and 4, 2 and 3, or both.



A web is defined to be interval if each predator can be represented by a line segment and the overlaps in the predators' diets are exactly represented by the overlaps among the line segments. The hypothetical graph above left is not interval because the interval that represents predator 3 cannot overlap the intervals that represent predators 1 and 4 without also overlapping the interval that represents predator 2, contrary to the graph.

For many arrangements of predator overlap graphs, the rigid circuit property ensures that the overlaps will have an interval representation. Yet the rigid circuit property does not guarantee that food webs are interval. Below is a non-interval, rigid-circuit predator overlap graph. There are no circuits around four or more species. It is still not possible to express the overlaps as overlapping segments of a line, because any segment for predator 1 that overlaps the segment for predator 2 must also overlap the segment for predator 3 or 4, contrary to the graph. This pattern is called *asteroidal* and the example comes from 7 species of gastropods in the genus *Conus*, from ref. 29 and original work by ref. 55.



Prey overlap graphs are formed by connecting prey species sharing one or more predators. Two prey sharing predators form a line, three a triangular plane, four a tetrahedron, and so on. Some prey species will be used by more than one predator and this allows us to assemble the individual graphs into a complete picture describing the relationships between predators and prey. Now consider a physical analogy of this overlap graph with the objects formed by connecting the various prey species that share a particular predator being considered to be solid. It is possible for this physical analogy to have holes; topological holes are their equivalent in the graph. Topological holes are rare in webs with small numbers of species⁸, but Fig. 4 provides an example.

hasty and uncritical comparisons of web features, especially now that the availability of a collection in machine readable format⁷ makes such comparisons easy. More convincing contrasts in patterns will be revealed within sets of webs specially collected for the purpose by the same scientists^{11,14-18,20}.

In web studies, omnivory has a special meaning (Box 1). Webs are said to contain omnivores when species can trace the energy they obtain through pathways of different lengths. Although still poorly studied critically by a single investigator, the extent of omnivory apparently varies between systems^{9,19}. Omnivory is statistically rare in some webs, but there are at least three important exceptions. Aquatic webs often have as their top predators fishes that start their lives as very small

zooplanktivores and thus 'eat their way up the food chain' as they grow. Detritivores often pay little attention to the former trophic position of species now dead. Finally, webs composed of insects and their parasitoids often contain species that feed facultatively on many different species at different trophic levels.

What causes the patterns?

How many independent web patterns are there? Which patterns are the consequences of others? What causes the patterns? A variety of evolutionary and natural history explanations have been put forward for particular web properties (reviewed in refs 9 and 25). Although we do not dismiss these explanations (most have been poorly investigated), we wish to focus on two more general explanations: the cascade model and web dynamics.

The cascade model

The cascade model⁶ focuses on the static patterns of trophic interaction and assigns linkages at random subject to two constraints. First, the model assumes that species can be arranged *a priori* into a cascade or hierarchy such that a given species can feed only on species below it, and itself can be fed on only by species above it in the hierarchy. This ordering automatically precludes cycles (Fig. 2; Boxes 1 and 3) and decomposer loops. It does not specify whether any particular species must be top, intermediate, or basal (except the lowest and highest species in the cascade). Second, the model requires two parameters obtained empirically: *S* and *d*. By assumption, connectance declines hyperbolically.

By assigning linkages randomly within these constraints, the cascade model generates quantitative predictions that can be compared rigorously with observed patterns. It correctly predicts the average and variance of the fractions of all species that are basal, intermediate, and top predators (Box 3); the average fractions of linkages that are basal-intermediate, basal-top, intermediate-intermediate and intermediate-top (Box 3); the modal length of chains from basal to top species (Box 3); and the decline in the frequencies of interval and rigid-circuit predator overlap graphs (Box 3) as webs get larger.

The cascade model has not yet been used to explore some features of webs that merit attention, such as omnivory, compartments (Box 3), and the ratios of how many prey species a species exploits to how many predatory species that species suffers (Box 3). It also gets some of the fine details wrong though, again, problems in the quality of the data may be partly responsible for the discrepancy. For example, predicted declines in interval and rigid-circuit predator overlap graphs as *S* increases are too rapid, and the predicted frequencies of very short and very long food chains within a given web are too high.

The assumption of constant linkage density is challenged by data on species-rich webs (above). An increase in linkage density with the number of species has been incorporated in a variant of the cascade model¹² which, surprisingly, produces a poorer fit to some other observed patterns than the original model. Obviously, the cascade model is very simple, making it easy to refine in various ways, while retaining the essential feature of a trophic cascade. Of 13 such variants¹² only one does as well as or, in some cases, better than the original model in predicting observed web properties. This variant has yet to be tested in detail.

The original formulation of the model offered no explanation for the postulated trophic cascade. Body size is a likely candidate²⁶ because typically predators are larger²⁷ than their prey, and parasites are smaller^{25,26}. The relationship between the size of a predator, the size of its prey, and their positions in the web is important²⁸ and not fully explored.

What determines the linkage density? Quantitative theory to explain why the average species apparently utilizes, and is utilized by, a predictable and fairly small number of other species is crucial to a deeper understanding and ultimately testing of the model. We offer one explanation below.

BOX 3. Features common to published community webs and their associated graphs

Features of the web itself

- Cycles are rare⁵. (The definitions of cycles and of other terms in this box are in Boxes 1 and 2.)
- The average proportion of top predators, intermediate species, and basal species remains roughly constant (but with high variance)⁵ in webs with widely differing numbers of species and from different habitats.
- The average proportion of trophic links that are between intermediate and intermediate species, intermediate species and top predators, basal species and intermediate species, and basal species and top predators remains constant⁵ (with large variance) in webs with widely differing numbers of species and from different habitats.
- Linkage density seems to be approximately constant for webs with few species, but may increase with large numbers of species (see text).
- For top predators, the modal food chain lengths are typically 2 or 3, with lengths of only 1 being less common (and perhaps representing incomplete information) and those greater than 3 being uncommon^{5,8}. Correspondingly, the modal number of trophic levels is 3 or 4.
- Omnivory is rare in some webs, though there are many exceptions (see text).
- Habitat boundaries sometimes impose compartments on food webs. Compartments are not usually observed within habitats⁹.
- For intermediate species, how do the numbers of species they exploit correlate with the numbers of species exploiting them? Studies of species at the same trophic levels in different systems (for instance insect herbivores or parasites) typically do not find a significant correlation^{9,54}. Within a food web, species at higher trophic levels tend to have more prey and fewer predators than species at low trophic levels, an observation predicted by the cascade model⁵.

Features of the predator overlap graphs

- Webs with small numbers of species only rarely have predator overlap graphs that are not rigid circuit⁵.
- Food webs with small numbers of species usually have interval overlaps in the predators' use of prey species⁵.
- For many arrangements of predator overlap graphs, the rigid circuit property ensures that the overlaps will be interval. Yet the rigid circuit property does not guarantee that food webs are interval because of the asteroidal pattern. Such asteroidal patterns of overlap are unusual in webs with small numbers of species⁵.

Features of the prey overlap graphs

- Topological holes are rare in webs with small numbers of species⁸.

Web dynamics

Most webs are static descriptions, but the communities they describe are not static. Some species are successful at invading a community whereas others are not and the successes may or may not cause extinctions of former residents. This process of assembly and disintegration may explain many of the empirical web patterns^{9,29}. It is not incompatible with the cascade model because it suggests general mechanisms that limit linkage density and species richness, while making specific predictions about the details of web patterns.

The idea that web patterns are shaped by dynamic constraints on assembly and disintegration was first formulated using Lotka-Volterra equations^{9,25}. (These are sets of quadratically nonlinear first-order constant coefficient differential equations. Each equation determines the rate of change of a species in the web, from the density of every species with which it interacts, as well as its intrinsic birth and death rates.) Such analyses distinguished stable (persistent) from unstable web structures and so concentrated on the disintegration of webs. In general, these models predict that systems with high linkage density *d* will be unstable; the critical value of *d* declines with increasing *S*. Recent models assembled complex communities from simpler ones and found comparable results: models with low *d* for a given *S* are more likely to be invaded³⁰⁻³². Moreover, once *d*

and S have come into dynamic equilibrium, each successful invasion makes subsequent invasions less likely^{32,33}.

Although based on extremely simple models, the processes uncovered are often intuitively sensible, as can be seen from their specific predictions. Consider the difficulties of entering highly connected webs. If invading predators overlap in their prey use with existing predators, levels of interspecific competition will increase, reducing the chance of successful invasion or making the persistence of competitors already in the web less likely. Either way, this process limits species richness and linkage density. Less familiar is the process of apparent competition³⁴ between species of prey that share the same predators. An increase in one prey species may cause an increase in the shared predator and a consequent decrease in the other prey species. High linkage density makes shared predators more likely, apparent competition stronger and webs harder to invade as a consequence. Competition and apparent competition acting in concert may generate constant predator-to-prey ratios in webs³⁵.

Species may enter the top of existing food chains. Long chains, however, reduce the rate at which population densities recover from environmental disasters⁹ making the persistence of species in long chains unlikely and invasion more difficult. This argument explains why chains appear shorter in unpredictable systems and conceivably explains why insect-dominated chains are short²². Insects have particularly variable population densities and may therefore have much larger minimum viable population densities than vertebrates²⁹.

The scarcity of omnivory in some systems may be due to the obvious difficulties encountered by a species that is both the prey and the competitor of a resident species. Alternatively, a successfully invading omnivore may locally exterminate a resident species on which it can both feed and with which it competes, so forming a food chain without omnivory. In contrast, when each of the omnivore's life history stages depends critically on resident species at progressively higher trophic levels as the omnivore matures, the omnivore cannot eliminate these intermediate species and still persist³⁶. Dynamic models also predict that omnivory will be common in decomposer systems (because the consumers are donor-controlled²⁵) and in communities of insects and their parasitoids²⁵. Other approaches to dynamic modelling of food webs are under investigation²⁸.

Concatenating dynamic and cascade models

Dynamic models contrast in purpose, technique, and accomplishment with the cascade model. The dynamic models attempt to explain why some web features are rarely or frequently observed in nature, without saying exactly how rarely or how frequently. These models rest on several largely untested assumptions³⁷ but offer insight into large questions of community structure, stability and change that are currently beyond the scope of the cascade model. By contrast, the cascade model aims for quantitative explanation, unification and prediction of observed patterns. It rests on extremely simple assumptions that are susceptible to direct test (possibly after additional interpretation, as when the cascade is supposed to be an ordering by body size).

Careful observations and experiments on spatial and temporal variation in webs may reveal more about the roles of the cascade and the dynamic models. For example, seasonal changes in web structure may differ from successional changes. Dynamic models predict that unstable webs from early successional habitats should depart significantly from the norm, for example, in having unusual and widely fluctuating predator-to-prey ratios³⁵. The predictions of the cascade model for this situation have yet to be developed.

In real ecological communities both population dynamics and trophic structure are important. A new hybrid model, the Lotka-Volterra cascade model (LVCM)³⁷, assumes the population dynamics of the Lotka-Volterra model when the interactions between species are shaped by a refinement of the cascade

model. The LVCM displays an ecological phase transition: gradual changes in the probabilities of various kinds of interactions related to feeding produce rapid changes in a community's probability of qualitative stability. The frontier of stability in this model differs from that of a classical model of May³⁸. The LVCM makes possible a unified discussion of stability and trophic structure in a single, analytically tractable framework, but inherits the shortcomings shared by its parents, the Lotka-Volterra equations and the cascade model.

Consequences of web patterns

The shape of food webs has inevitable dynamic consequences for constituent species. Our understanding of these consequences is growing quickly^{29,39}. For example, the rate at which populations recover from disasters (resilience) theoretically should depend on food chain length⁹, an idea now supported by a limited number of experimental studies⁴⁰. Variability of population density depends on features of the species examined and on the variability of the physical environment. It also depends on web features such as the number of prey the species exploits, and the number of predatory species that exploit it^{41,42}.

Removal of individuals is a common experimental procedure in ecology⁴³⁻⁴⁵. Food web theory warns that it may be difficult to predict responses of unmanipulated species. For instance, permanent removal of a species' predator may ultimately cause the species' density to decline if the predator also fed on a powerful competitor. Theory predicts that even when the web is known, slight changes in the parameters that describe interspecific interactions may be amplified or dampened by complex pathways generating indirect effects between any pair of species⁴⁶⁻⁴⁸. In contrast to these final consequences of removals, there should be little doubt about the direction of the transient changes⁴⁶. Thus, prey species should initially increase when their predators are removed: it takes time for indirect effects to appear. When we do see changes in density expected from immediate interactions with prey and predators, it may only be because of the typically short duration of field studies. At intermediate time scales, we should see effects rippling through the web, a feature illustrated by the relatively long-term studies of the removals of ants and rodents from desert communities⁴⁹.

Web theory has addressed changes in species composition, and in the total density of a group of species when species are manipulated or removed²⁹. Theory predicts that highly-connected (complex) communities should be most sensitive to the loss of species from the top of the web because secondary extinctions propagate more widely than in loosely connected (simple) communities²⁹. In contrast, simple communities should be more sensitive to the loss of plant species than complex

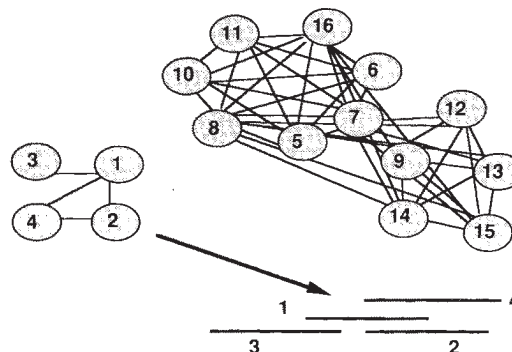


FIG. 3 Predator overlap graph for the *Nepenthes* web. In the *Nepenthes* web there are two connected sets of predators one involving four predators, the other involving the other twelve. The overlap graph in the component of the *Nepenthes* graph involving predators 1 through 4 is rigid circuit. Species 3 in the figure does not violate this condition because it is not part of a circuit around four or more species. In the *Nepenthes* web, the observed predator overlap involving species 1 to 4 is interval—as can be seen from the overlaps of the segments above right. (See Box 2 for terminology).

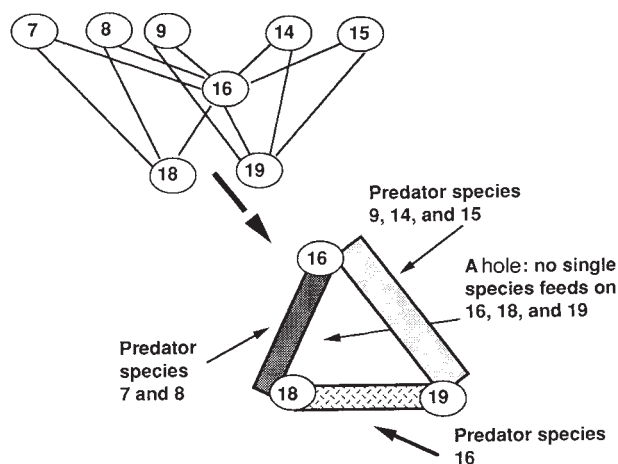


FIG. 4 Prey overlap graph derived from the *Nepenthes* web and the 'topological hole' formed by connecting the various prey species that share a particular predator (see Box 2).

communities, because in simple communities the consumers are dependent on only a few species and cannot survive their loss²⁹. 'Keystone' species, those having disproportionate effects when removed, can occur at all trophic levels. Theory⁵⁰ has also addressed empirical findings⁵¹ that when grazers are removed from a community, increased plant diversity is associated with large changes in species composition, but relatively small changes in total plant density.

Whither food webs?

A major obstacle to progress in food web research is the weakness of the available data; this is more severe than a matter of not having enough webs. A bigger problem is the lack of methodological standards; the aim must be to reduce inter-observer subjective differences in the way in which field workers observe and report webs. To assess the completeness of observations, Cohen⁵ recommended a yield-effort curve, which plots the cumulative number of species and linkages against sampling effort. In a recent example⁵², there is little sign of a decrease in

the number of new species observed per field day after more than 5,000 hours of observation over 10 years. This finding supports the criticism, often made, that many published webs are grossly incomplete in the species and perforce in the linkages included. In such detailed studies, however, many of the linkages must represent rare and thus possibly dynamically unimportant events. The need to quantify the relative frequency of feeding interactions is crucial.

Despite methodological and theoretical problems, we favour a cautiously optimistic view of the future of food web research. Present shortcomings in theory and data must serve as a spur to refining both; they are not a reason to discard emerging insights and understanding.

Nature presents ecologists with ecological communities as the natural units of analysis. Though boundaries may be drawn in different ways for different purposes, the given primitives of ecology are forests, lakes, watersheds, wetlands, shorelines, deserts, estuaries, oceans, and the like. Populations of a single species, and still more individuals of a single population, are abstractions from communities for the sake of analytical convenience. To a considerable degree, these abstractions have upstaged the communities from which ecologists isolate them. In this time of rapid environmental change, the many endangered species are only a drop in the flood of endangered communities. Although it is by no means the only, or a sufficient approach, a quantitative, predictive understanding of food webs would provide a better basis for solving many problems of applied ecology than is now available. The average proportions of species in different trophic positions is found to be roughly invariant in webs with different numbers of species: this finding has already been used, for example, in an estimate of the total number of species on earth from the number of known plant species⁵³. We anticipate increasing use of food web theory in such disparate but pressing problems as the management of multispecies fisheries, integrated pest control, and predicting the effects of climate change on ecological communities. □

Stuart L. Pimm is at the Department of Zoology and Graduate Program in Ecology, University of Tennessee, Knoxville, Tennessee 37996, USA, and at the NERC Centre for Population Biology, Imperial College, Silwood Park, Ascot, Berks SL5 7PY, UK, John H. Lawton is at the NERC Centre for Population Biology, and Joel E. Cohen is at the Rockefeller University, New York, New York 10021, USA.

1. Darwin, C. *On the Origin of Species by Means of Natural Selection* (John Murray, London, 1859). (Facsimile reprint: Harvard University Press, Cambridge, MA, 1964).
2. Forbes, S. A. *Ecological Investigations of Stephen Alfred Forbes (1844-1932)* (reprint: Arno Press, New York, 1977).
3. Pierce, W. D., Cushman, R. A. & Hood, C. E. *U.S. Dept. Agric. Bull.* **100**, 9-99 (1912).
4. Shelford, V. E. *Bull. Geogr. Soc. Chicago* **5** (1913) (reprint: Arno Press, New York, 1977).
5. Cohen, J. E. *Food Webs and Niche Space* (Princeton University Press, Princeton, NJ 1978).
6. Cohen, J. E., Briand, F. & Newman, C. M. *Community food webs: Data and theory* (Springer-Verlag, New York, 1990).
7. Cohen, J. E. *Ecologists Co-operative Web Bank (ECOWEB™). Version 1.0. Machine Readable Data Base of Food Webs* (Rockefeller University, New York, 1989).
8. Sugihara, G. *Proc. Symp. appl. Math.* **30**, 3-101 (1984).
9. Pimm, S. L. *Food Webs* (Chapman and Hall, London, 1982).
10. Paine, R. T. *Ecology* **69**, 1648-1654 (1988).
11. Jeffries, J. J. & Lawton, J. H. *Freshwater Biol.* **15**, 105-112 (1985).
12. Cohen, J. E. *Theor. Pop. Biol.* **37**, 55-90 (1990).
13. Sugihara, G., Schoenly, K. & Trombka, A. *Science* **245**, 48-52 (1989).
14. Warren, P. *Am. Nat.* **316**, 689-700 (1990).
15. Winemiller, K. O. *Am. Nat.* **134**, 960-968 (1989).
16. Beaver, R. A. *Ecol. Ent.* **10**, 241-248 (1985).
17. Kitching, R. L. & Pimm, S. L. *Proc. Ecol. Soc. Austr.* **14**, 123-139 (1986).
18. Kitching, R. L. & Beaver, R. A. in *Living in Patchy Environments* (eds Shorrocks, B. & Swingland, I. R.) 147-175 (Oxford University Press, Oxford, 1990).
19. Schoener, T. W. *Ecology* **70**, 1559-1589 (1989).
20. Kitching, R. L. *Oikos* **48**, 280-288 (1987).
21. Schoenly, K. & Cohen, J. E. *Ecol. Monogr.* (in the press).
22. Schoenly, K., Beaver, R. A. & Heurnier, T. A. *Am. Nat.* (in the press).
23. Yodzis, P. *Oecologia (Berl.)* **65**, 86-88 (1984).
24. Pimm, S. L. & Kitching, R. L. *Oikos* **50**, 302-307 (1987).
25. Lawton, J. H. in *Ecological Concepts* (ed. Cherrett, J. M.) 43-78 (Blackwell Scientific Publications, Oxford, 1989).
26. Warren, P. H. & Lawton, J. H. *Oecologia (Berl.)* **74**, 231-235 (1987).
27. Vézina, A. F. *Oecologia (Berl.)* **67**, 555-565 (1985).
28. Carpenter, S. R. *Complex Interactions in Lake Communities* (Springer-Verlag, New York, 1988).
29. Pimm, S. L. *The Balance of Nature?* (University of Chicago Press, Chicago, 1991).

30. Robinson, J. V. & Valentine, W. D. *J. theor. Biol.* **81**, 91-104 (1979).
31. Shigesada, N., Kawasaki, K. & Teramoto, E. *Theor. Pop. Biol.* **36**, 311-338 (1989).
32. Post, W. M. & Pimm, S. L. *Math. Biosci.* **64**, 169-192 (1983).
33. Drake, J. A. *J. theor. Biol.* **147**, 213-234 (1990).
34. Holt, R. D. *Theor. Pop. Biol.* **12**, 197-229 (1977).
35. Mithen, S. J. & Lawton, J. H. *Oecologia (Berl.)* **69**, 542-550 (1986).
36. Pimm, S. L. & Rice, J. A. *Theor. Pop. Biol.* **32**, 303-325 (1987).
37. Cohen, J. E., Luczak, T., Newman, C. M. & Zhou, Z. M. *Proc. R. Soc. B* **240**, 607-627 (1990).
38. May, R. M. *Nature* **238**, 413-414 (1972).
39. Pimm, S. L. *Nature* **307**, 321-326 (1984).
40. Redfearn, A. & Pimm, S. L. in *Insect Pests* (eds Barbosa, P. & Schultz, J. C.) 99-133 (Academic Press, Orlando, FL, 1987).
41. Redfearn, A. & Pimm, S. L. *Ecol. Monogr.* **58**, 39-55 (1988).
42. Redfearn, A. & Pimm, S. L. in *Predation* (ed. Crawley, M.) (Blackwell Scientific Publications, Oxford, 1991).
43. Schoener, T. W. *Am. Nat.* **122**, 240-285 (1983).
44. Connell, J. H. *Am. Nat.* **122**, 661-696 (1983).
45. Sih, A., Crowley, P., McPeck, M., Petranka, J. & Strohmeier, K. A. *Rev. Ecol. Syst.* **16**, 269-311 (1985).
46. Yodzis, P. *Ecology* **69**, 508-515 (1988).
47. Patten, B. C. *Can. J. Fish. Aqu. Sci.* **213**, 119-138 (1985).
48. Lane, P. A. in *The Biology of Mutualism: Ecology and Evolution* (ed. Boucher, D. H.) 345-374 (Croom Helm, London, 1985).
49. Brown, J. H., Davidson, D. W., Munger, J. C. & Inouye, R. S. in *Community Ecology* (eds Diamond, J. & Case, T. J.) 41-62 (Harper & Row, New York, 1986).
50. King, A. W. & Pimm, S. L. *Am. Nat.* **122**, 229-239 (1983).
51. McNaughton, S. J. *Ecol. Monogr.* **55**, 259-294 (1985).
52. Polis, G. *Am. Nat.* (in the press).
53. May, R. M. *Science* **241**, 1441-1449 (1988).
54. Hawkins, B. A. & Lawton, J. H. *Nature* **326**, 788-790 (1987).
55. Kohn, A. J. *Ecol. Monogr.* **29**, 47-90 (1959).

ACKNOWLEDGEMENTS. Authors are listed in reverse alphabetical order. We thank P. Warren for helpful suggestions. J.E.C. was partly supported by the NSF and by the hospitality of Mr and Mrs W. T. Golden. During the preparation of this paper S.L.P. was supported by the NERC Centre for Population Biology and the University of Tennessee; he also thanks J. K. Pimm.

Medium-range order in the cation distribution of a calcium silicate glass

P. H. Gaskell*, M. C. Eckersley*, A. C. Barnes^{†‡} & P. Chieux[†]

* University of Cambridge, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK

[†] Institut Laue-Langevin, Avenue des Martyrs, 38042 Grenoble Cedex, France

Silicate glasses have conventionally been regarded as silicate frameworks in which cations are distributed at random. Neutron scattering from isotopically substituted samples allows correlations between cations to be investigated beyond the nearest-neighbour coordination shell, and shows that a degree of ordering persists over distances approaching 1 nm. The structural picture that emerges has elements of long-range randomness coexisting with both short- and medium-range order.

DEFINITION of the nature and extent of atomic ordering in simple amorphous solids is an important experimental and theoretical challenge. The structure of many diatomic oxide glasses, such as B_2O_3 , SiO_2 and GeO_2 , measured over a distance scale greater than ~ 2 nm is best represented as a disordered, isotropic solid but with considerable order on a scale of nearest-neighbour interatomic distances. For the alkali and alkaline-earth silicate glasses, the conventional (Zachariasen-Warren) model^{1,2} suggests that the structure consists of a framework formed by SiO_4 tetrahedra, into the interstices of which cations such as calcium are generally thought to fit. These interstices were thought to be poorly defined, but it has recently become clear^{3,4} that the first-neighbour shell is relatively ordered. The distribution of these 'network-modifying' atoms is, however, still an open question, a random distribution determined solely by macroscopic parameters such as composition and density being the general assumption. Here we present direct evidence that the distribution of calcium atoms in a calcium silicate glass is non-random, with features that reflect a more extensive structural organization bearing many of the characteristics of the crystalline form.

For the oxide glasses, typified by the silicate described here, the experimental challenge is to establish atomic distribution functions with sufficient precision and information content to generate and test structural models. In earlier work⁴, we used neutron scattering with isotopic substitution of Ca in a silicate glass to obtain a weighted sum of the subset of partial pair-correlation functions centred on the calcium atom. This technique revealed a high degree of ordering in the immediate environment of Ca. Here we extend the technique to obtain a direct measurement of the Ca-Ca distribution, which provides strong evidence that such glasses are more extensively ordered than previously seemed possible.

The glass chosen for this investigation is identical to that studied earlier⁴: $(CaO)_{48.0}(SiO_2)_{49.0}(Al_2O_3)_{3.0}$. This is close to the metasilicate composition, allowing comparison with data for crystalline $CaSiO_3$ (wollastonite). It is free from sub-liquidus immiscibility, as shown by high-resolution transmission electron micrographs⁵, and crystallization is suppressed by the addition of a small amount of Al_2O_3 . We took care in melting and

quenching to ensure that only the isotopic concentration of Ca differed in three specimens containing essentially pure natural Ca (^{nat}Ca), ^{44}Ca and an equal mixture of the two (^{mix}Ca). The scattered intensities were measured using the twin-axis diffractometer D4 at the Institut Laue-Langevin, Grenoble. We then obtained two first differences between the scattered intensities, $I(nat) - I(mix)$ and $I(mix) - I(44)$.

The normalized scattering intensity is proportional to the differential scattering cross-section, $d\sigma/d\Omega$, corrected for absorption, multiple scattering and inelasticity:

$$\frac{1}{N} \frac{d\sigma}{d\Omega} = \frac{\sigma_s}{4\pi} + \sum_{\alpha\beta} c_\alpha c_\beta b_\alpha b_\beta (S_{\alpha\beta}(Q) - 1) \quad (1)$$

Here N is the number of scattering atoms, σ_s is the sum of coherent and incoherent cross-sections, and c_α and b_α are atomic fractions and scattering lengths for atoms of species α . $S_{\alpha\beta}(Q)$ is a Faber-Ziman partial structure factor which contains information on the arrangement of β atoms around α , and $Q = 4\pi \sin \theta / \lambda$, where 2θ is the scattering angle and λ the neutron wavelength.

Subtraction of differential cross-sections for two isotopically substituted specimens cancels identical terms in $S_{\alpha\beta}(Q)$ for atom pairs that exclude Ca. Separating terms involving like atoms leads to:

$$\begin{aligned} \Delta_{Ca}(Q) &= c_{Ca}^2 (b_{Ca}^2 - b_{Ca}''^2) (S_{CaCa}(Q) - 1) \\ &+ 2 \sum_{\beta \neq Ca} c_{Ca} c_\beta b_\beta \Delta b (S_{Ca\beta}(Q) - 1) \end{aligned} \quad (2)$$

Here, $\Delta b = b_{Ca} - b_{Ca}''$, where b_{Ca} and b_{Ca}'' are the scattering lengths for the two Ca isotopes and differences in the Q -independent term $\sigma_s/4\pi$ have been subtracted.

Previously⁴ we reported a measurement of the first difference, $\Delta_{Ca}(Q)$, and analysis of the weighted pair correlation function $G_{Ca}(r)$ obtained by Fourier transformation:

$$\begin{aligned} G_{Ca}(r) &= 2/\pi \int_0^\infty Q \Delta_{Ca}(Q) \sin(Qr) dQ \\ &= c_{Ca}^2 (b_{Ca}^2 - b_{Ca}''^2) G_{CaCa}(r) \\ &+ 2 \sum_{\beta \neq Ca} c_{Ca} c_\beta b_\beta \Delta b G_{Ca\beta}(r) \end{aligned} \quad (3)$$

The functions $G_{\alpha\beta}(r)$ are defined by

$$G_{\alpha\beta}(r) = 4\pi r \rho_0 (g_{\alpha\beta}(r) - 1) \quad (4)$$

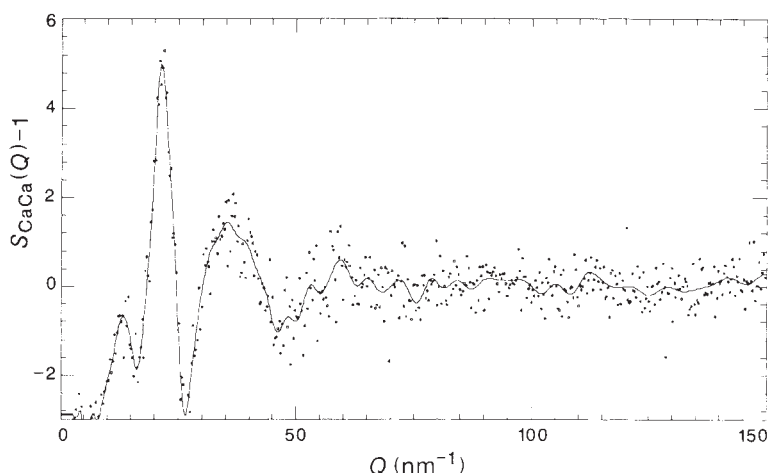
with $g_{\alpha\beta}(r) = 1 + (2\pi^2 r \rho_0)^{-1} \int_0^\infty Q (S_{\alpha\beta}(Q) - 1) \sin(Qr) dQ$, and ρ_0 the average atomic density.

For the glass considered here, $G_{Ca}(r)$ is dominated by the Ca-O distribution, especially at low values of r . The Ca-O first-neighbour distances centred on 0.237 nm can be extracted with considerable precision: about 85% of the 6.15 (± 0.17) oxygen neighbours lie in a narrow distribution with a standard deviation of 0.012 nm, compared with an estimated 0.011 nm for the crystal. The remaining (0.95) oxygen atoms occupy a broad tail stretching from ~ 0.25 to 0.285 nm.

In $c\text{-}CaSiO_3$, Ca-Ca and Ca-Si distances are almost equal (0.36 and 0.34 nm respectively). In the glass, a pronounced second peak is seen at 0.35 nm, and the earlier indirect evidence

[‡] Present address: H. H. Willis Physics Laboratory, University of Bristol, Royal Fort, Tyndall Avenue, Bristol BS8 1TL, UK.

FIG. 1 Experimental double-difference data $S_{\text{CaCa}}(Q)$ for the calcium silicate glass. Points are raw data; full line is a smoothed version obtained by fitting the differential cross-sections with a cubic spline before subtraction. The signal is significantly greater than the noise level only for $Q \leq 80 \text{ nm}^{-1}$.



suggests that this feature does, in part, represent Ca-Ca first-neighbour distances. An alternative assumption of a random distribution of Ca atoms in the Si-O network leads to a larger mean separation, 0.44 nm.

Although this circumstantial evidence favours a Ca-Ca distance associated with a bonding scheme similar to that of the crystal, experimental first-difference data alone cannot discriminate unambiguously between the two possibilities. Nor can it be claimed that the first coordination sphere of about six oxygens is in an octahedral arrangement, and nothing can be said about the manner in which Ca-centred polyhedra are linked. In c-CaSiO₃, octahedra are linked at edges forming infinite trioctahedral Ca-O sheets. A direct measurement of the Ca-Ca pair correlation function is thus required to establish, firmly, whether randomness or a degree of medium-range ordering should form the pattern for future structural modelling of this and similar materials.

The isotopic concentrations of the three glasses were chosen such that Δb values corresponding to the first-difference data sets were identical. Subtraction of the two first differences then cancels terms other than those involving $S_{\text{CaCa}}(Q)$. Thus:

$$\Delta(\Delta_{\text{Ca}}(Q)) = c_{\text{Ca}}^2 [\Delta_1(b_{\text{Ca}}^2) - \Delta_2(b_{\text{Ca}}^2)] [S_{\text{CaCa}}(Q) - 1] \quad (5)$$

where $\Delta_1(b_{\text{Ca}}^2)$ and $\Delta_2(b_{\text{Ca}}^2)$ are differences in the squares of the scattering lengths for each pair of Ca isotopes.

Simplifying the expression for the double difference leads to:

$$2\Delta(\Delta_{\text{Ca}}(Q)) = c_{\text{Ca}}^2 (b_{\text{Ca}} - b_{\text{Ca}}'')^2 (S_{\text{CaCa}}(Q) - 1) \quad \text{or} \quad (6)$$

$$S_{\text{CaCa}}(Q) = 1 + 2\Delta(\Delta_{\text{Ca}}(Q)) / (c_{\text{Ca}}^2 (b_{\text{Ca}} - b_{\text{Ca}}'')^2)$$

where b and b'' are the scattering lengths for the two end members, and the scattering length of the intermediate composition, b_{Ca}' , is chosen so that $b_{\text{Ca}} - b_{\text{Ca}}' = b_{\text{Ca}}' - b_{\text{Ca}}''$.

A double-difference experiment thus gives direct information on the Ca-Ca distribution. The experimental problem, however, is that the signal is very weak and subject to sizeable systematic and random errors. The 'total' structural signal oscillates around the isotropic term $\sigma_s/4\pi$ ($\sim 25 \text{ fm}^2$) with an amplitude $\sum_{\alpha\beta} c_{\alpha} c_{\beta} b_{\alpha} b_{\beta} (S_{\alpha\beta}(Q) - 1)$, which is typically one-fifth of this value. Realistic counting times (~ 1 day for a 5-g specimen) limit the signal-to-noise ratio to $\sim 0.2\%$ for the total function. The double-difference signal, of order $c_{\text{Ca}}^2 (b - b'')^2 / 2 = 0.18 \text{ fm}^2$, thus carries a random noise of $\sim 30\%$. Alternatively expressed, the double-difference signal is over 600 times smaller than $S_{\text{CaCa}}(Q)$. Systematic errors are even more serious. Errors in composition of $\sim 0.5\%$ lead to 'contamination' of the double-difference data by the much stronger Si-O, O-O and Ca-O correlations. Because specimen masses are limited to ~ 6 g by the expense of the ^{44}Ca isotope, reproducible melting to this level of accuracy is not guaranteed, and this is the chief limitation to accuracy at present.

Figure 1 shows $S_{\text{CaCa}}(Q)$, both as the original data and as a function obtained by subtraction of cubic spline fits to $1/N d\sigma/d\Omega$. Some error in these data, ascribed to inadequate subtraction of contributions other than Ca-Ca pairs, is immediately evident in oscillations of $S_{\text{CaCa}}(Q) - 1$ below the limit of -1.0 . (It would not be unreasonable to normalize the data to the limiting value.) The Fourier transforms of $Q(S_{\text{CaCa}}(Q) - 1) = G_{\text{CaCa}}(r)$ for both sets of data are essentially identical, and results for the smoothed data are shown in Fig. 2. The first peak in $G_{\text{CaCa}}(r)$ at 0.23 nm is presumably due to incomplete removal of the first-difference signal. Examination of Fourier transforms truncated at different values of Q_{max} shows that a large contribution to the feature at 0.23 nm originates from oscillations in $S_{\text{CaCa}}(Q)$ beyond 80 nm^{-1} where the double-difference signal is vanishingly small. The other main features of $G_{\text{CaCa}}(r)$ are insensitive to Q_{max} , apart from inevitable loss of resolution. The most important feature is a strong, well-defined peak at ~ 0.36 – 0.39 nm. If this is the first Ca-Ca distance, R_1 , then correspondence with first-neighbour Ca-Ca distances in c-CaSiO₃ at 0.36 nm is excellent. Together with the indirect evidence obtained from first-difference data⁴, this direct measurement indicates first Ca-Ca distances close to the values observed in the crystalline phase. There is little support for a mean Ca-Ca distance near 0.44 nm as predicted by a random model.

The Ca-Ca coordination number is subject to considerable (and as yet unquantified) errors, but a fit to the data suggests that ~ 5.2 neighbours lie between 0.36 and 0.40 nm, with a further atom in a broad shell at ~ 0.46 nm. In c-CaSiO₃, there are 4.67

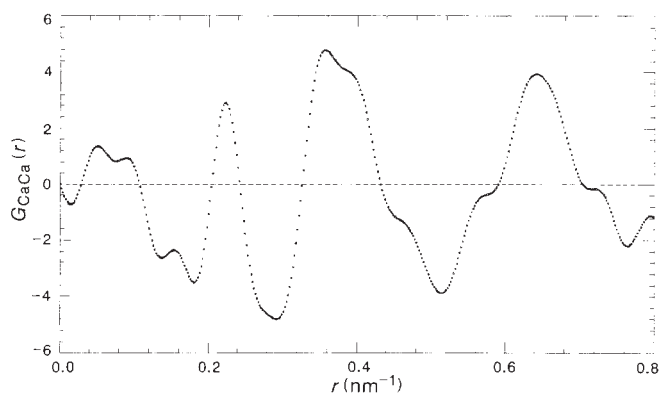


FIG. 2 Fourier transform of the double-difference data, $G_{\text{CaCa}}(r)$, obtained from smoothed $Q(S_{\text{CaCa}}(Q) - 1)$ data truncated at $Q_{\text{max}} = 110 \text{ nm}^{-1}$. Differences in $G_{\text{CaCa}}(r)$ introduced by smoothing are small. The strong peak at ~ 0.38 nm probably corresponds to the nearest Ca-Ca distance and is close to values found in c-CaSiO₃. The strong peak at 0.64 nm probably corresponds to the second Ca-Ca neighbours. The first peak at ~ 0.23 nm is thought to arise from inadequate subtraction of the first-difference signals (see text).

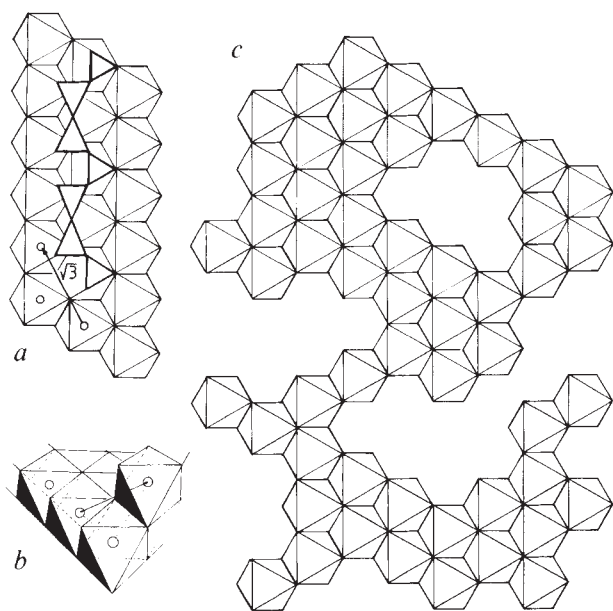


FIG. 3 *a*, Schematic view of trioctahedral strips of edge-sharing CaO_6 octahedra in c-CaSiO_3 . A polymeric $(\text{SiO}_3)_\infty$ chain is shown between planes of close-packed oxygen anions. The arrow marks a Ca-Ca distance equal to $\sqrt{3}R_1$, where R_1 is the nearest-neighbour Ca-Ca distance. *b*, Corrugation of the sheet of edge-shared octahedra introduces distances such as that shown, in which the Ca-Ca distance (arrowed) is $\sqrt{2}R_1$. *c*, Part of a random sheet of edge-shared CaO_6 octahedra. The relative number of second and third neighbours (and so on) is related to the fractality of the sheet.

neighbouring Ca atoms in a broad distribution between 0.35 and 0.37 nm.

Features in the Ca-Ca distribution beyond 0.4 nm offer information on which new structural models can be based. One possibility, that the Ca-Ca distribution can be modelled in terms of dense-packed spheres, was considered and rejected for data on barium-containing glasses⁶. The idea was revived more recently⁷ where peaks assigned to Cs-Cs pairs in caesium silicate glasses were shown to lie close to ratios of first, second and higher neighbour distances obtained in dense-packed models for monoatomic metals. Here we follow an alternative approach based on suggestions made previously^{8,9,10}. The structures of crystalline pyroxenes, and pyroxenoids such as wollastonite, can be considered as a close-packed oxygen sub-lattice with cations such as Ca occupying octahedral interstices in alternate layers parallel to $\{111\}$ (in a face-centred cubic model). Layers of Si atoms are interleaved, the silicons occupying tetrahedral interstices (Fig. 3*a*). It is arguable that this structural principle persists in glasses. A more controversial idea is that the need for polyvalent cations such as Ca to satisfy coordination requirements might be a dominant constraint, with Si-O chains suitably adapting their conformations to fit.

We propose as a working hypothesis that pyroxene and pyroxenoid glasses can be represented by distinguishable domains which are best described in terms of close-packing of oxygens. Sheets of interconnected CaO_6 octahedra are essentially parallel in these regions and are interleaved with linked SiO_4 tetrahedra. Defects can be described as corrugation and 'tearing' of the sheets, and variations in connectivity in the planes. In addition to such topological disorder, there will be geometrical distortions introduced by the topological 'defects'. Figure 3 shows these schematically.

The broad second peak centred on 0.64 nm observed in $G_{\text{CaCa}}(r)$ coincides with $\sqrt{3}R_1 = 0.65$ nm. The significance of this is that close-packed $\{111\}$ planes of CaO_6 octahedra have second-neighbour Ca-Ca distances of $\sqrt{3}R_1$ (Fig. 3*a*). This additional information suggests that the 6.15 oxygen neighbours of Ca, for which it was impossible to assign any symmetry from first-difference data alone, form Ca-centred octahedra which are

edge-linked in glasses as well as in crystals.

Almost as revealing as the presence of peaks in $G_{\text{CaCa}}(r)$ are their absences. A model based on randomly corrugated sheets of CaO_6 octahedra would involve a significant number of $\sqrt{2}R_1$ distances at 0.53 nm arising from the out-of-plane corrugations (Fig. 3*b*). This distance corresponds to a minimum in the experimental $G_{\text{CaCa}}(r)$ data, however, so this otherwise tempting model seems to be ruled out. Calcium-oxygen sheets must be essentially planar—topologically, if not geometrically.

A second significant absence is noted at $2R_1 = 0.75$ nm. A broad shoulder is seen in $G_{\text{CaCa}}(r)$ at this distance but is quite insignificant: the relative weakness of this feature may be related to disorder in the Ca-O sheet. Although on average this must be a trioctahedral sheet (valence and composition forbid significant deviations), disorder in the glass could be represented by random occupancy of octahedral interstices in a planar net. The ratio of $\sqrt{3}R_1$ and $2R_1$ distances then increases from 1.0 in the crystal to ~ 1.7 in the bifurcated sheet structure (Fig. 3*c*). The relative magnitude of the peaks at $\sqrt{3}R_1$ and $2R_1$ is thus a crude measure of the fractality of the Ca-O sheet, and the experimental data indicate large local deviations from the tri-octahedrality.

Finally, the weaker correlations corresponding to distances in c-CaSiO_3 between Ca atoms separated by planes of Si-O tetrahedra, are also absent. These occur at distances ranging from 0.5 to 0.6 nm in the crystal, corresponding to a prominent minimum in the experimental curve for the glass. Again, fractal sheets of Ca and O atoms suggest the answer. Unless defects in adjacent sheets are correlated, which seems unlikely, interplanar distances cannot be well-defined.

The new information offered by this double-difference measurement has strengthened the evidence for ordered structures in silicate glasses, and possibly in other amorphous oxides. The indication from first-difference measurements that a Ca-Ca first-neighbour distance of 0.38 nm corresponds to a value typical of crystals is confirmed, and the peak in $G_{\text{CaCa}}(r)$ at 0.64 nm also indicates that the cation environment is octahedral, the octahedrons being edge-linked. The extent of sheet formation, and of deviations from this, can also be assessed from the absence of peaks in $G_{\text{CaCa}}(r)$ corresponding to the interplanar separations $\sqrt{2}R_1$ and $2R_1$. All these features are consistent with the theory that Ca lies in regions where the structure, although distorted, approximates to cubic close-packing. It would be difficult otherwise to understand the well defined local structure around a weakly bonded ion. As argued previously for metallic glasses¹¹, the presence of a well-defined local structure around such elements has wide implications for medium-range organization: In close-packed glasses (or in the close-packed regions of silicates and other oxide glasses) the occurrence of local ordering around highly-interconnected atoms such as Ca is a consequence of a well-developed medium-range structure. The evidence presented here goes some way to explore this territory. The double-difference technique has possible applications in structural studies of liquids, melts and disordered solids. Our more recent measurements include mixed Ca-Mg and Ca-Ni silicate and copper phosphate glasses. □

Received 22 November 1990; accepted 15 March 1991.

1. Zachariasen, W. H. *J. Am. chem. Soc.* **54**, 3841–3851 (1932).
2. Warren, B. E. & Bischoff, J. *J. Am. Ceram. Soc.* **21**, 259–265 (1938).
3. Greaves, G. N. in *Glass Science and Technology* **4B** (eds Uhlmann, D. R. & Kreidl, N. J.) 1–76 (Academic, London, 1990).
4. Eckersley, M. C., Gaskell, P. H., Barnes, A. C. & Chieux, P. *Nature* **335**, 525–527 (1988).
5. Saeed, A. thesis, Univ. of Cambridge (1990).
6. Block, S. & Piermarini, G. J. *Phys. Chem. Glasses* **5**, 138–143 (1964).
7. Hanson, C. D. & Egami, T. *J. Non-cryst. Solids* **87**, 171–184 (1986).
8. Dent-Glasser, L. S. *Z. Kristallogr.* **149**, 291–305 (1979).
9. Liebau, F. *Structural Chemistry of Silicates* (Springer, Berlin, 1985).
10. Law, A. D. & Whittaker, E. J. W. *Mineralog. Mag.* **43**, 565–574 (1980).
11. Gaskell, P. H. *J. Non-cryst. Solids* **75**, 329–346 (1988).

ACKNOWLEDGEMENTS. We thank Pilkington plc (P.H.G.) and SERC (M.C.E.) for financial support, and P. Boden, D. R. Codrham, W. H. Flynn, H. L. McPhail, S. J. Monelly, G. W. F. Pardoe and B. Yale for help in glass melting and advice.

High-affinity NGF binding requires coexpression of the *trk* proto-oncogene and the low-affinity NGF receptor

Barbara L. Hempstead*, Dionisio Martin-Zanca^{†‡}, David R. Kaplan[§], Luis F. Parada^{*} & Moses V. Chao^{*}

* Hematology/Oncology Division, Department of Medicine, Department of Cell Biology and Anatomy, Cornell University Medical College, 1300 York Avenue, New York, New York 10021, USA

† Molecular Embryology Group, § Eukaryotic Signal Transduction Group, ABL-Basic Research Program, NCI-Frederick Cancer Research & Development Center, PO Box B, Frederick, Maryland 21701, USA

Nerve growth factor (NGF) interacts with two different low-affinity receptors that can be distinguished by affinity crosslinking. Reconstitution experiments by membrane fusion and transient transfection into heterologous cells indicate that high-affinity NGF binding requires coexpression and binding to both the low-affinity NGF receptor and the tyrosine kinase *trk* gene product. These studies reveal a new growth factor receptor-mediated mechanism of cellular differentiation involving *trk* and the low-affinity NGF receptor.

CELL survival and differentiation in the nervous system are mediated by numerous growth factors, of which NGF has provided the most compelling evidence for a physiological role during development¹⁻³. NGF is a selective neurotrophic factor for sympathetic and some sensory neurons of the peripheral nervous system, and for the ascending cholinergic neurons of the basal forebrain. Furthermore, administration of exogenous NGF can rescue developing neurons that would otherwise undergo naturally occurring cell death.

A key question regarding the action of NGF in responsive cells has been the molecular mechanism of signal transduction after ligand binding to the functional NGF receptor. The many effects of NGF on selective neuronal populations are directly dependent on initial binding of NGF to specific cell-surface receptors. Like other polypeptide growth factors, NGF interacts with responsive cells with biphasic equilibrium binding kinetics, reflecting two classes of receptors. Approximately 10–15% of the receptor sites display high-affinity binding (dissociation constant, $K_d = 10^{-11}$ M) and the remainder display low-affinity binding ($K_d = 10^{-9}$ M)⁴⁻⁶. The concentration of NGF necessary to elicit a response⁷ and the detection of high-affinity NGF binding only in cells responsive to NGF⁸⁻¹⁰ indicate that biological responsiveness to NGF depends on interactions with the high-affinity form of the receptor. No biological responses mediated solely by the low-affinity receptor have been observed^{11,12}. These differences imply that the molecular events mediating the change in receptor affinity are involved in the signal transduction of NGF.

Complementary DNAs representing the NGF receptor messenger RNA have been isolated after gene transfer of genomic DNA and screening with monoclonal antibodies that either block or alter NGF binding in human melanoma cells¹³ or rat PC12 cells¹⁴. The sequence of the cDNA for human¹⁵, rat¹⁶ and chicken NGF receptor^{17,18} indicates that the receptor

is a transmembrane protein of relative molecular mass 42,000 (M_r 42K), which is glycosylated to yield a protein of about 75K *in vitro* (p75^{NGFR}).

Although the cysteine-rich extracellular region of p75^{NGFR} shares homology with the B-cell-activation molecule, CDw40 (ref. 19), the tumour-necrosis-factor receptor²⁰⁻²², and T-cell-activation molecule, OX40 (ref. 23), no significant homology exists for the cytoplasmic region, including any similarity to tyrosine or serine/threonine kinases. When expressed in many cell lines, p75^{NGFR} displays no functional responses following ligand binding¹², and generates only low-affinity NGF-binding sites with a K_d of 10^{-9} M (refs 16, 24). Generation of both high- and low-affinity sites is observed only when NGF receptor cDNAs is introduced into selective cell lines of neuronal origin²⁵⁻²⁷. These observations suggest that the low-affinity NGF receptor participates in the formation of the high-affinity receptor complex, but that other cell-specific molecules are required for the generation of the high-affinity receptor and for transmembrane signalling²⁵.

NGF has been shown to bind to the proto-oncogene *trk*, and induce autophosphorylation by the *trk*-encoded tyrosine kinase in PC12 cells^{28,29}. The product of the proto-oncogene *trk* is a transmembrane glycoprotein that is selectively expressed in the developing nervous system. The identification of the proto-oncogene *trk* (p140^{prototrkr}) as an NGF receptor in PC12 cells suggests that p140^{prototrkr} could participate in high-affinity NGF binding and be involved in mediating the specific actions of NGF. Here we demonstrate that p140^{prototrkr} displays a different interaction with NGF than does p75^{NGFR}, although both receptors bind with a similar low-affinity K_d . Furthermore, the coexpression of proto-oncogene p140^{prototrkr} and p75^{NGFR} are required for the formation of high-affinity NGF-binding sites.

Crosslinking of NGF receptors

Affinity crosslinking of NGF-responsive cells with ¹²⁵I-labelled NGF has been extensively used to characterize receptors for NGF. When the hydrophilic reagent ethyldimethylaminopropyl carbodiimide (EDAC) is used, only one complex of M_r 90–100K is seen (Fig. 1). This complex has been observed in whole brain preparations^{30,31}, and a variety of cell types including Schwann cells, melanoma, pheochromocytoma and neuroblastoma cells³²⁻³⁵. The complex can also be immunoprecipitated with antibodies against NGF, as well as with monoclonal antibodies against the receptor; the complex therefore consists of NGF and the low-affinity NGF receptor, p75^{NGFR}. Other crosslinking reagents, such as the photoactivatable agent *N*-hydroxysuccinimidyl-4-azidobenzoate (HSAB), generate two crosslinked complexes of 100K and 158K (ref. 36). The 158K crosslinked complex is generated only in cells that have high-affinity binding sites, including pheochromocytoma cells and neurons. A variety of models have been proposed to address the relationship between the two crosslinked species, including one in which the

† Present address: Instituto de Microbiología Bioquímica, CSIC, Universidad de Salamanca, 37008 Salamanca, Spain.

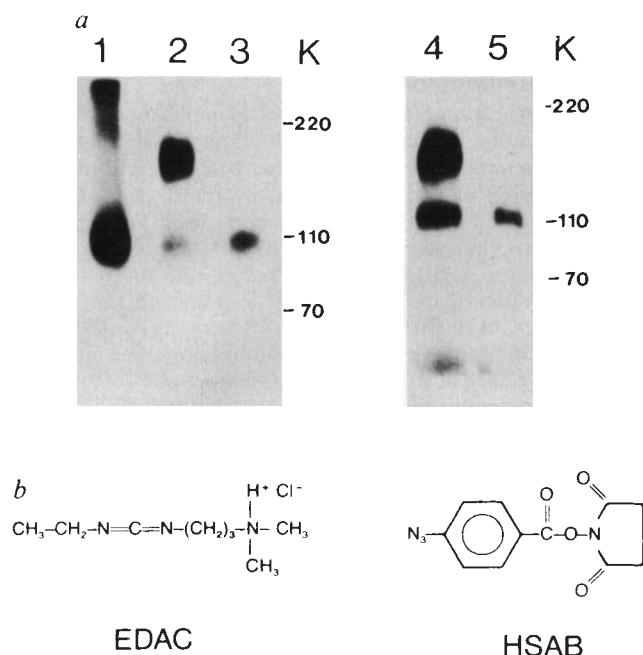


FIG. 1 *a*, Affinity crosslinking of ^{125}I -labelled NGF to pheochromocytoma PC12 cells. PC12 cells (2×10^6 per ml) were incubated with 0.5 nM ^{125}I -labelled NGF for 2 h at 4°C and then crosslinked using EDAC²⁵ (lane 1) or HSAB^{31,47} (lanes 2–5). Detergent extracts were immunoprecipitated as described⁴⁷ using anti-NGF antisera (Collaborative) (lanes 1, 2 and 4), IgG-192¹⁴ antisera (lane 3), or anti-p75^{NGFR} antibody specific for amino acids 249–265 (lane 5). Immunoprecipitates were analysed by SDS-PAGE and autoradiography. About 5×10^6 cells were analysed in lane 1, and 30×10^6 cells were used in lanes 2–5. Exposure, 2 days at -70°C . *b*, Structure of crosslinking reagents, EDAC and HSAB (Pierce).

158K species consists of a complex of p75^{NGFR} and a 60K protein^{11,37}.

Recent HSAB crosslinking has firmly identified the 158K species as p140^{prototrkr} crosslinked to NGF²⁹. This complex can be coimmunoprecipitated with p75^{NGFR} in PC12 cells with antibodies to mouse NGF (Fig. 1). However, three independently derived antibodies directed against different epitopes of p75^{NGFR} do not recognize the 158K crosslinked species. These antibodies are: the IgG-192 monoclonal antibody¹⁴; an anti-peptide antibody specific for amino acids 249–265 (Fig. 1*a*, lane 5); an antibody specific for amino acids 163–396 of rat p75^{NGFR} (data not shown), a region encompassing 60 amino acids of the extracellular domain, the transmembrane domain and all of the cytoplasmic region. These experiments confirm that two distinct protein species can bind NGF, and support earlier biochemical

data describing the different proteolytic susceptibility, pH sensitivity and Triton X-100 solubility of the two kinetic forms of the receptor^{6,38–40}.

An explanation for why the efficiency of crosslinking of p140^{prototrkr} to ^{125}I -labelled NGF is different with EDAC than it is with HSAB (Fig. 1*b*) could lie in the mechanism of action of the crosslinking agents. EDAC is hydrophilic and directly crosslinks proteins by using a free carboxylic acid group; HSAB is hydrophobic, interacts with amino groups, and introduces an 8 Å spacer arm between the crosslinked proteins. Because EDAC does not crosslink NGF with p140^{prototrkr} on PC12 cells, different protein–protein contacts are probably made between NGF and each of the two receptors. Indeed, the extracellular sequences of the human *trk* proto-oncogene product⁴¹ and human p75 receptor¹⁵ have very little amino-acid similarity, particularly in the pattern of their cysteine residues.

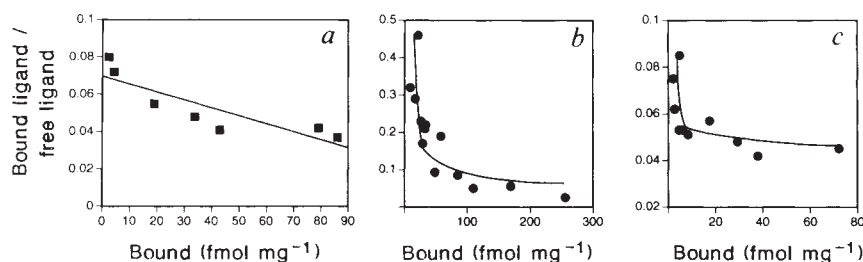
These crosslinking experiments illustrate fundamental differences between the interaction of NGF with p75^{NGFR} and its interaction with p140^{prototrkr}. The two species defined by HSAB crosslinking also differ in their phosphorylation, with the 100K species primarily having phosphorylated serine and threonine residues^{42,43} and the 158K species containing phosphotyrosine^{28,44}. Further, the high- and low-affinity forms of the receptor seem to differ in their cytoskeletal interactions^{6,38} and in their ability to undergo ligand mediated internalization^{37,45,46}.

Membrane fusion experiments

Previous experiments using biochemical crosslinking of p75^{NGFR} to ^{125}I -labelled NGF demonstrated that the 100K crosslinked species is common to both the high- and low-affinity forms of the NGF receptor³⁴. This model, in which p75^{NGFR} plus additional cellular components is required to produce the high-affinity form of the receptor, has been supported by gene transfer experiments. In these studies, a cDNA encoding the human p75^{NGFR} was introduced into cell lines of neuronal origin that lack the receptor. The resulting transfectants displayed both high- and low-affinity sites^{25,27}, the 100K and 158K crosslinked species⁴⁷, and responded to NGF^{25–27}. Because p140^{prototrkr} crosslinked to NGF is the 158K species, we tested whether p140^{prototrkr} is directly involved in the conversion of low-affinity to high-affinity NGF binding by performing membrane fusion experiments.

A 3T3 transformant line overexpressing rat p140^{prototrkr} (rtkr-3T3) binds ^{125}I -labelled NGF with low affinity²⁹. Membranes isolated from this line were fused using polyethylene glycol (PEG)⁴⁸ to membranes from a fibroblast line (PAE1c) expressing p75^{NGFR}, and the final membrane suspension was incubated with increasing concentrations of ^{125}I -labelled NGF. As previously demonstrated, PAE1c cells express receptors with only low-affinity binding sites²⁵. Similarly, membranes isolated from rtkr-3T3 cells had an equilibrium binding constant consistent with low-affinity binding²⁹. However, fusion of membranes from

FIG. 2 ^{125}I -labelled NGF equilibrium binding analysis of fused membranes. Crude membrane preparations were isolated as described²⁵ from PAE1c and PAB9 cells (PA317 fibroblasts expressing p75^{NGFR})^{25,47}, mouse 3T3 fibroblasts, rtkr-3T3 (ref. 29) and NR18 cells⁵¹. Binding data were plotted according to Scatchard⁴⁹. *a*, PAE1c membranes (1.32 mg) fused with 3T3 membranes (2.3 mg). *b*, PAE1c membranes (4 mg) fused with rtkr-3T3 membranes (0.40 mg). *c*, NR18 membranes (2.25 mg) fused with PAB9 membranes (2.2 mg). METHODS. Membrane preparations were fused using PEG according to Robb⁴⁸, and equilibrium binding analysed by vacuum filtration as described²⁵. Each binding reaction was carried out in triplicate, with and without a thousand-fold excess of unlabelled NGF. Binding data was corrected for nonspecific binding by subtracting values obtained in the presence of unlabelled NGF. Specific binding



ranged from 60 to 90% of total binding. Amount of membrane protein used for each reaction was 30 µg. ^{125}I -labelled NGF was prepared by lactoperoxidase treatment.

PAE1c cells and rtk-3T3 cells produced both high- and low-affinity ^{125}I -labelled NGF binding by Scatchard⁴⁹ analysis (Fig. 2b). Using the LIGAND program⁵⁰, two site analysis and curve fitting generated equilibrium dissociation constants of 2.5×10^{-11} M and 1.4×10^{-9} M, consistent with values obtained in ^{125}I -labelled NGF binding experiments with sensory neurons and cultured PC12 cells^{4,6}. The generation of the biphasic curve was the direct consequence of fusion of p140^{prototrkr}-containing membranes with PAE1c membranes, because fusion of membranes from untransfected 3T3 cells with PAE1c membranes (Fig. 2a) produced a linear Scatchard plot ($K_d = 1.5 \times 10^{-9}$ M), reflecting binding to low-affinity NGF receptors in PAE1c cells.

This experiment supports the previous gene transfer experiments suggesting that additional cell-specific molecules are required for high-affinity binding. One of those experiments used the pheochromocytoma cell line NR18 that lacks p75^{NGFR} as the recipient cell. These cells have been described as incapable of binding NGF⁵¹, implying that p140^{prototrkr} is not expressed in them. But RNA blot analysis indicates that NR18 cells express *trk* messenger RNA, but at very low levels compared with PC12 cells (Fig. 3). In addition, HSAB crosslinking reveals extremely low amounts of p140^{prototrkr} in NR18 cells (data not shown).

To verify that NR18 cells can reconstitute high-affinity binding, PEG-mediated membrane fusion experiments were carried out with fibroblast cells expressing p75^{NGFR} and NR18 cells. NR18 membranes did not detectably bind NGF, as determined by filter binding, presumably owing to the low numbers of p140^{prototrkr} molecules in this mutant cell line. However, fusion of NR18 membranes with membranes containing p75^{NGFR} (Fig. 2c) yielded both high- and low-affinity classes of receptors ($K_d = 2.5 \times 10^{-11}$ M and 1.4×10^{-9} M). The much lower number of high-affinity sites (300 sites per cell) reconstituted after membrane fusion is consistent with the low level of *trk* expression in this mutant PC12 cell line. These experiments also indicate that in the absence of p140^{prototrkr} receptors (Fig. 2a), no high-affinity binding could be detected. Furthermore, fusion of membranes with PEG is necessary to generate the biphasic Scatchard

plot, because omission of PEG during the procedure leads to only low-affinity binding (data not shown).

Transient transfection of *trk*

The membrane fusion experiments directly implicate p140^{prototrkr} as the protein that participates in high-affinity NGF binding in membrane preparations. To test whether p140^{prototrkr} is involved in the formation of high-affinity sites by transfection, COS-1 cells and human melanoma cells were used in transient expression experiments. The full-length cDNA encoding rat *trk* and a 1.5-kilobase (kb) cDNA representing the full length of the coding region of the human p75^{NGFR} were subcloned into the plasmid pCMV5 vector⁵².

COS cells were transfected with each plasmid alone and also with both together at a ratio of 1 to 10 (*trk*: p75^{NGFR}). COS cells do not normally express *trk* mRNA or p140^{prototrkr} (Figs 3 and 4), but 48 hours after transfection, abundant *trk* mRNA and p140^{prototrkr} are produced. The appearance of the receptors at the cell surface and their ability to interact with NGF was verified by affinity crosslinking. HSAB crosslinking of ^{125}I -labelled NGF to COS cells transfected with the *trk* cDNA yields a 158K species, whereas affinity crosslinking of COS cells expressing p75^{NGFR} yields a 90K crosslinked species. When the two plasmids were used in transient transfectants, both the 158K and 90K crosslinked species were generated (Fig. 4, lane 7). Crosslinked p140^{prototrkr} was also recognized specifically using anti-Trk antibodies (lane 8), whereas untransfected COS cells did not display any crosslinked species when either reagent was used. The relative lack of crosslinking of p140^{prototrkr} by EDAC (Fig. 4, lane 5) confirms that the interaction of NGF with p140^{prototrkr} in PC12 cells is recapitulated in transfected COS cells.

Crude membranes were isolated from transfected COS cells and assayed by filter binding using ^{125}I -labelled NGF. Membranes prepared from cells expressing p140^{prototrkr} alone generated a linear Scatchard plot (Fig. 5a), with a K_d of 3.3×10^{-9} M, consistent with the equilibrium binding constant measured in

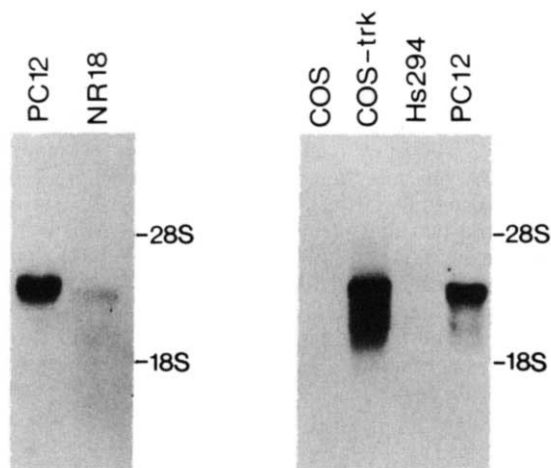


FIG. 3 RNA analysis of *trk* mRNA expression. Total RNA isolated from PC12 (40 μg), NR18 (40 μg), Hs294 (20 μg) and COS cells (20 μg). COS-*trk* RNA (20 μg) was generated by transient transfection of a calcium phosphate precipitate⁵³ containing the expression plasmid pDM115 (10 μg per dish) for the rat *trk* cDNA. After overnight incubation, the cells were treated with 20% glycerol for 1 min. The slightly longer mRNA for the transfected *trk* arises because of an upstream start site in the expression plasmid. Migration positions of ribosomal RNA markers are shown on the right. METHODS. RNA was isolated from confluent cells and separated by electrophoresis in a 1% agarose gel containing 2.2 M formaldehyde. Expression of *trk* mRNA was assessed by hybridization to a 464-base-pair insert from pDM97, specific for rat *trk*. The probe was radiolabelled by primer extension and hybridized at 50% formamide, $5 \times \text{SSC}$, $1 \times \text{Denhardt's}$ solution, salmon sperm DNA (100 $\mu\text{g ml}^{-1}$) and 10% dextran sulphate. Filters were washed to $0.5 \times \text{SSC}$ at 60°C followed by autoradiography.

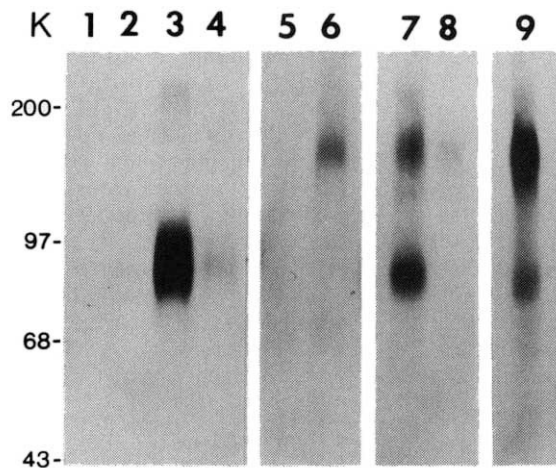
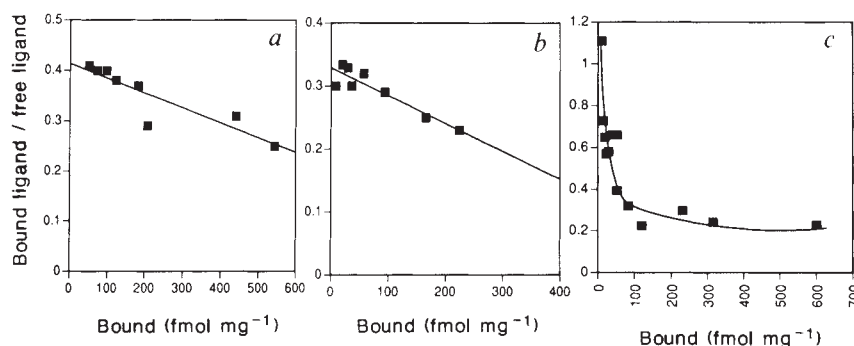


FIG. 4 Affinity crosslinking of p140^{prototrkr} and p75^{NGFR} expressed after transient transfection in COS cells using pCMV-*trk* and pCMV5A (p75^{NGFR})¹⁵. Lane 1, COS cells (1 ml), crosslinked with EDAC; lane 2, COS cells (1 ml) crosslinked with HSAB; lane 3, COS cells (2 ml) transfected with pCMV5A, crosslinked with EDAC; lane 4, COS cells (2 ml) transfected with pCMV5A, crosslinked with HSAB; lane 5, COS cells (2 ml) transfected with pCMV-*trk*, crosslinked with EDAC; lane 6, COS cells (2 ml) transfected with pCMV-*trk*, crosslinked with HSAB; lane 7, COS cells (4.5 ml) transfected with 4 μg pCMV-*trk* and 10 μg pCMV5A per dish, crosslinked with HSAB; lane 8, same as lane 7 except immunoprecipitation was carried out with anti-Trk antisera; and lane 9, PC12 cells (1 ml) crosslinked with HSAB. METHODS. Cells (2×10^6 per ml) were incubated with 1 nM ^{125}I -labelled NGF for 2 h at 4°C , followed by crosslinking as described previously⁴⁷. After detergent lysis, extracts were immunoprecipitated with anti-NGF antisera (lanes 1–7 and 9) or anti-p140^{prototrkr} (lane 8), washed extensively, and subjected to SDS-PAGE.

FIG. 5 Equilibrium binding analysis of transfected receptors expressed in COS cells. COS cells were transiently transfected, and collected after 60 h for crude membrane preparation and binding with ^{125}I -labelled NGF. Equilibrium binding analysis was performed as described in Fig. 2. *a*, COS cells transfected with pDM115 (p140^{prototrkr}). *b*, COS cells transfected with pCMV5A (p75^{NGFR}). *c*, COS cells transfected with pDM115 and pCMV5A at a 1 to 10 ratio.



membranes from stably transfected 3T3 cells²⁹. Although these results do not exclude the presence of a few higher-affinity binding sites, computer analysis of many different COS cell transfectants and fibroblast lines have yielded equivalent K_d values. COS cells expressing p75^{NGFR} also generated a linear Scatchard plot, with a K_d of 1.1×10^{-9} M (Fig. 5*b*). These equilibrium binding constants are consistent with earlier studies of p75^{NGFR} expression in heterologous cells^{16,24,25}. Immunofluorescence of the transfected cells indicates that at least 30–40% of the COS cells express p75^{NGFR}. When membrane preparations from cells transfected with both *trk* and NGF receptor cDNAs were examined, both high- and low-affinity sites with a K_d of 3×10^{-11} M and 5×10^{-9} M, respectively, were found, in close agreement with the values obtained by PEG-mediated fusion of fibroblast membranes. Because these gene transfer experiments used the same type of cells and the same transfection protocol and equilibrium binding conditions, expression of both cDNAs appears to be required for high-affinity binding.

To confirm that coexpression of the p75^{NGFR} and the p140^{prototrkr} receptors are required to reconstitute high-affinity binding sites, the *trk* cDNA was transiently expressed in the human Hs294 melanoma cell line, which expresses many p75 receptors⁵⁴, but no *trk* mRNA (Fig. 3). Hs294 cells express p75^{NGFR} at the cell surface and display only low-affinity sites with a K_d of 5.9×10^{-9} M (Fig. 6*a*). Addition of NGF to these cells does not result in any of the responses associated with NGF treatment of PC12 cells. But when Hs294 cells were transfected with the *trk* expression plasmid pDM115, binding to ^{125}I -labelled NGF yielded a curvilinear Scatchard plot, indicating a high-affinity site with a K_d of 3.1×10^{-11} M and a lower-affinity site with a K_d of 5.5×10^{-9} M (Fig. 6*b*). Hence, low-affinity NGF binding can be converted to high-affinity binding with the expression of *trk* in transfected melanoma cells.

Discussion and implications

The molecular basis for high- and low-affinity NGF receptors has been an enigma. The results reported here provide strong support for the hypothesis that the high-affinity NGF binding site is composed of at least two NGF-binding components: the *trk* gene product and p75^{NGFR}, each of which binds NGF with

a low-affinity K_d . This conclusion is supported by three experiments: (1) membrane fusion studies using cells independently expressing p75^{NGFR} and p140^{prototrkr}; (2) gene transfer experiments in which the p75^{NGFR} and p140^{prototrkr} receptors are coexpressed in a cell line that normally expresses neither; and (3) gene transfer of a *trk* cDNA in cells of neuronal origin which have high levels of endogenous expression of p75^{NGFR}.

Previous investigations of the molecular characteristics of the high-affinity and low-affinity receptor classes have not clearly defined the two species. Initial studies used affinity crosslinking of radiolabelled NGF to cells which express both high- and low-affinity receptors. Using HSAB⁵⁵, two complexes of 158K and 100K were detected³⁶. In contrast, HSAB crosslinking of cells that express only the low-affinity NGF receptor yields only the 100K crosslinked species. These results led to the conclusion that the 158K species represents the high-affinity receptor crosslinked to NGF, and that the 100K species represents the low-affinity receptor crosslinked to NGF. Further evidence demonstrating that the two species differ in their trypsin sensitivity, Triton X-100 solubility, dissociation rate^{5,6}, pattern of phosphorylation^{28,42–44}, and lack of crossreactivity in immunoprecipitation experiments, strongly suggested that the two binding sites are represented by two independent protein moieties.

Other biochemical studies, however, suggested that the high-affinity receptor had features in common with the low-affinity receptor. When affinity crosslinking was performed using EDAC, only the 100K crosslinked complex was seen in PC12 cells bearing both high- and low-affinity receptors over a wide range of NGF concentration³⁴. Gene transfer experiments further supported the idea that p75^{NGFR} participates in the formation of the high-affinity receptor complex. Introduction of p75^{NGFR} into appropriate cell lines resulted in the reconstitution of both high- and low-affinity sites, responsiveness to NGF treatment^{25,26,56}, and the appropriate HSAB crosslinking pattern (both 158K and 100K species)⁴⁷. These results suggest that in addition to the low-affinity NGF receptor, another protein is required for formation of the high-affinity site.

The results of the study described here indicate that both p140^{prototrkr} and p75^{NGFR} are required to form the high-affinity ligand-binding site. Membrane fusion studies imply that p75^{NGFR} and p140^{prototrkr} receptor proteins do not need to be synthesized concomitantly to form high-affinity binding sites. In addition, reconstitution of high-affinity ligand binding by cotransfection of COS cells suggests that the appropriate translation and post-translational modification of the two proteins, and their migration to the cell surface to form the high-affinity site, do not necessarily require a neuronal cell environment.

Although signal transduction by NGF is likely to be initiated by tyrosine phosphorylation by p140^{prototrkr}, the precise mechanisms that regulate ligand affinity remain unclear. Given the requirement for both p75^{NGFR} and p140^{prototrkr}, a model of high affinity receptor site formation can be proposed (Fig. 7). From affinity-crosslinking and equilibrium-binding studies, it is known that both proteins can interact with NGF, although it is uncertain whether they interact with a monomer or dimer of NGF. Analysis of equilibrium binding of ^{125}I -labelled NGF to p140^{prototrkr} clearly argues against p140^{prototrkr} alone being the high-affinity

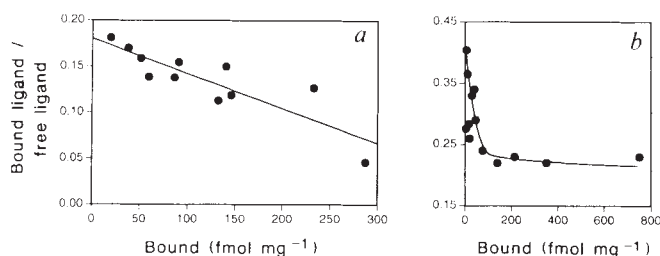


FIG. 6 Equilibrium binding analysis of HS294 melanoma cells expressing p140^{prototrkr}. Hs294 melanoma cells⁵⁴ were transiently transfected with pDM115 using calcium phosphate⁵³, glycerol-shocked and collected for membrane preparation 60 h later. NGF binding was performed on membranes by filter binding²⁵. Membranes were isolated from untransfected Hs294 cells (*a*) and Hs294 cells transfected with pDM115 (*b*).

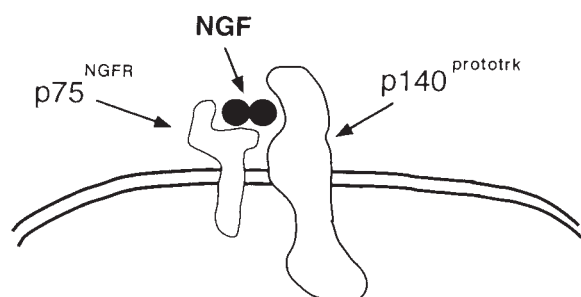


FIG. 7 Hypothetical model for high-affinity NGF binding. NGF exists as a noncovalent β dimer that can bind both to $p75^{NGFR}$ and $p140^{prototrkr}$. Previous experiments indicated that $p140^{prototrkr}$ alone can bind specifically to NGF²⁹. Reconstitution of high-affinity binding after membrane fusion and transfection of *trk* in heterologous cells supports a model in which both $p75^{NGFR}$ and $p140^{prototrkr}$ are required.

NGF receptor, when $p140^{prototrkr}$ is expressed independently of the low-affinity NGF receptor.

The binding studies do not address the role of $p75^{NGFR}$ in directly modulating signal transduction. However, in previous studies, cells expressing mutant $p75^{NGFR}$ lacking portions of the highly conserved $p75$ cytoplasmic domain fail to generate high-affinity sites when expressed in NR18 cells⁴⁷. In addition, NR18 cells expressing these mutants lack normal tyrosine phosphorylation of cellular proteins in response to NGF⁵⁷. These studies suggest that the cytoplasmic domain of $p75^{NGFR}$ may interact with $p140^{prototrkr}$, and that this interaction leads not only to the generation of a high-affinity site, but also to appropriate transmembrane signalling, as detected by tyrosine phosphorylation. Further experiments with chimaeric receptors containing cytoplasmic portions of the $p75$ molecule also support the role of $p75^{NGFR}$ in signal transduction by NGF⁵⁸.

Our reconstitution experiments do not exclude the possibility of additional components in the fully functional high-affinity NGF receptor. Among the examples of multisubunit receptors are the interleukin-2 (IL-2) receptor with at least two subunits (α and β) directly interacting with IL-2 (ref. 59), and four other proteins associated with the receptor complex⁶⁰⁻⁶², and the interleukin-6 (IL-6) receptor, with an IL-6 binding subunit and a nonbinding associated protein, gp130, which acts in signal transduction⁶³. In addition, heparin-like molecules are required for high-affinity binding of basic fibroblast growth factor to its receptor⁶⁴.

Intracellular molecules may also couple with the high-affinity NGF receptor complex after ligand activation. As has been previously demonstrated with the receptors for platelet-derived growth factor^{65,66} and epidermal growth factor⁶⁷, and for the CD4 and CD8 antigens⁶⁸, the cytoplasmic domains of these transmembrane receptors can interact with a variety of soluble enzymes, including GAP, phospholipase C- γ and $p56^{lck}$, when the appropriate receptor is activated after ligand binding. From the model presented above, such protein-protein interactions could occur with the *trk* protein alone, $p75^{NGFR}$ alone, or both species together.

Elucidation of the structural components of the functional high-affinity receptor is not only crucial in defining the mechanism of action of NGF, but will provide insight into the mode of signal transduction of other related neurotrophic factors. Brain derived neurotrophic factor (BDNF)⁶⁹ and neurotrophin-3 (NT-3) share significant homology with NGF^{70,71} and act in a similar manner *in vitro*, enhancing neuronal survival and promoting cellular differentiation⁷²⁻⁷⁵. BDNF and NT-3 differ from NGF in that discrete subpopulations of neurons that are unresponsive to NGF, such as placode-derived nodose sensory neurons, respond to BDNF and NT-3 (refs 72, 76). Also, the developmental regulation and sites of synthesis of BDNF and NT-3 are distinct from those described for NGF^{77,78}.

BDNF binds to sensory neurons in a saturable and specific

manner to high- and low-affinity sites⁷⁹. Furthermore, BDNF can recognize and bind to $p75^{NGFR}$ expressed in fibroblast cells, with kinetics similar to those of low-affinity binding⁸⁰. NT-3 appears to bind to PC12 cells, but does not induce a differentiated phenotype⁷⁵, suggesting that the high-affinity NGF receptor complex can discriminate between NT-3 and its appropriate ligand NGF. Indeed, $p75^{NGFR}$ could bind many different members of the neurotrophin family, and therefore can be considered to be a neurotrophin receptor. Because related *trk* tyrosine kinases such as *trkB* are expressed in distinct regions of the nervous system⁸¹⁻⁸³, specificity of neurotrophic action is most probably generated by the expression of an appropriate *trk* tyrosine kinase. The identification of $p140^{prototrkr}$ as an NGF receptor that participates as a subunit of the high-affinity NGF-binding site will not only further our knowledge about the mechanism of signal transduction for NGF, but will also serve as a model receptor system for discriminating the actions of many different neurotrophic factors. □

Received 11 March; accepted 4 April 1991.

- Levi-Montalcini, R. & Angeletti, P. U. *Physiol. Rev.* **48**, 534-569 (1968).
- Thoenen, H. & Barde, Y.-A. *Physiol. Rev.* **60**, 1284-1335 (1980).
- Levi-Montalcini, R. *Science* **237**, 1154-1164 (1987).
- Sutter, A., Riopelle, R. J., Harris-Warrick, R. M. & Shooter, E. M. *J. biol. Chem.* **254**, 5972-5982 (1979).
- Landreth, G. E. & Shooter, E. M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 4751-4755 (1980).
- Schechter, A. L. & Bothwell, M. A. *Cell* **24**, 867-874 (1981).
- Greene, L. A. *Brain Res.* **133**, 350-353 (1977).
- Sonnenfeld, K. H. & Ishii, D. N. *J. Neurosci. Res.* **8**, 375-391 (1982).
- Sonnenfeld, K. H. & Ishii, D. N. *J. Neurosci.* **5**, 1717-1728 (1985).
- Green, S. H., Rydel, R. E., Connolly, J. L. & Greene, L. A. *J. Cell Biol.* **102**, 830-843 (1986).
- Johnson, E. M. & Taniuchi, M. *Biochem. Pharmacol.* **36**, 4189-4195 (1987).
- Hempstead, B. L., Patil, N., Olson, K. & Chao, M. *Cold Spring Harb. Symp. quant. Biol.* **53**, 477-485 (1988).
- Ross, A. H. et al. *Proc. natn. Acad. Sci. U.S.A.* **81**, 6681-6685 (1984).
- Chandler, C. E., Parsons, L. M., Hosang, M. & Shooter, E. M. *J. biol. Chem.* **259**, 6882-6889 (1984).
- Johnson, D. et al. *Cell* **47**, 545-554 (1986).
- Radeke, M. J. et al. *Nature* **325**, 593-597 (1987).
- Large, T. H. et al. *Neuron* **2**, 1123-1134 (1989).
- Heuer, J. G., Fatemie-Nainie, S., Wheeler, E. F. & Bothwell, M. *Dev. Biol.* **137**, 287-304 (1990).
- Stamenkovic, I., Clark, E. A. & Seed, B. *EMBO J.* **8**, 1403-1410 (1989).
- Loetscher, H. et al. *Cell* **61**, 352-359 (1990).
- Schall, T. J. et al. *Cell* **61**, 361-370 (1990).
- Smith, C. A. et al. *Science* **248**, 1019-1023 (1990).
- Mallet, S., Fossum, S. & Barclay, A. N. *EMBO J.* **9**, 1063-1068 (1990).
- Chao, M. V. et al. *Science* **232**, 418-421 (1986).
- Hempstead, B. L., Schieffer, L. S. & Chao, M. V. *Science* **243**, 373-375 (1989).
- Matsushima, H. & Bogenmann, E. *Molec. cell. Biol.* **10**, 5015-5020 (1990).
- Pleasure, S. J. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8496-8500 (1990).
- Kaplan, D. R., Martin-Zanca, D. & Parada, L. F. *Nature* **350**, 158-160 (1991).
- Kaplan, D. R., Hempstead, B. L., Martin-Zanca, D., Chao, M. V. & Parada, L. F. *Science* (in the press).
- Taniuchi, M., Schweitzer, J. B. & Johnson, E. M. *Proc. natn. Acad. Sci. U.S.A.* **83**, 1950-1954 (1986).
- Escandon, E. & Chao, M. V. *Dev. Biol.* **142**, 293-300 (1990).
- Grob, P. M., Berlot, C. H. & Bothwell, M. A. *Proc. natn. Acad. Sci. U.S.A.* **80**, 6819-6823 (1983).
- Taniuchi, M., Clark, H. B. & Johnson, E. M. *Proc. natn. Acad. Sci. U.S.A.* **83**, 4094-4098 (1986).
- Green, S. H. & Greene, L. A. *J. biol. Chem.* **261**, 15316-15326 (1986).
- Hempstead, B. L. & Chao, M. V. *Recent Prog. Horm. Res.* **45**, 441-466 (1989).
- Hosang, M. & Shooter, E. M. *J. biol. Chem.* **260**, 655-662 (1985).
- Hosang, M. & Shooter, E. M. *EMBO J.* **6**, 1197-1202 (1987).
- Vale, R. D. & Shooter, E. M. *J. Cell Biol.* **94**, 710-717 (1982).
- Block, T. & Bothwell, M. J. *Neurochem.* **40**, 1654-1663 (1983).
- Buxser, S. E., Watson, L. & Johnson, G. L. *J. cell Biochem.* **22**, 219-233 (1983).
- Martin-Zanca, D. et al. *Molec. cell. Biol.* **9**, 24-33 (1989).
- Grob, P. M., Ross, A. H., Koprowski, H. & Bothwell, M. A. *J. biol. Chem.* **260**, 8044-8049 (1985).
- Taniuchi, M., Johnson, E. M., Roach, P. J. & Lawrence, J. C. *J. biol. Chem.* **261**, 13342-13349 (1986).
- Meakin, S. O. & Shooter, E. M. *Neuron* **6**, 153-163 (1991).
- Olender, E. J. & Stach, R. W. *J. biol. Chem.* **255**, 9338-9343 (1980).
- Bernd, P. & Greene, L. A. *J. biol. Chem.* **259**, 15509-15516 (1984).
- Hempstead, B. L., Patil, N., Thiel, B. & Chao, M. V. *J. biol. Chem.* **265**, 9595-9598 (1990).
- Robb, R. J. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3992-3996 (1986).
- Scatchard, G. *Ann. N.Y. Acad. Sci.* **52**, 660-672 (1949).
- Munson, P. J. & Rodbard, D. *Analyt. Biochem.* **107**, 220-239 (1980).
- Bothwell, M. A., Schechter, A. L. & Vaughn, K. M. *Cell* **21**, 857-866 (1980).
- Andersson, S. et al. *J. biol. Chem.* **264**, 8222-8231 (1989).
- Wigler, M. et al. *Cell* **16**, 777-785 (1979).
- Fabricant, R. F., Delarco, J. E. & Todaro, G. J. *Proc. natn. Acad. Sci. U.S.A.* **74**, 565-569 (1977).
- Massague, J. et al. *J. biol. Chem.* **256**, 9419-9424 (1981).
- Volante, C. & Greene, L. A. *Growth Factors* **2**, 321-331 (1990).
- Berg, M. M., Sternberg, D., Hempstead, B. L. & Chao, M. V. (manuscript submitted).
- Yan, H., Schlessinger, J. & Chao, M. V. *Science* (in the press).
- Hatakeyama, M. et al. *Science* **244**, 551-556 (1989).
- Waldeman, T. A. *Ann. Rev. Biochem.* **58**, 875-911 (1989).
- Sharon, M., Gnarra, J. R. & Leonard, W. J. *Proc. natn. Acad. Sci. U.S.A.* **87**, 4869-4873 (1990).
- Burton, J. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 7329-7333 (1990).
- Hibi, M. et al. *Cell* **63**, 1149-1157 (1990).
- Yayon, A., Klagsbrun, M., Esko, J. D., Leder, P. & Ornitz, D. M. *Cell* **64**, 841-848 (1991).
- Morrison, D. K., Kaplan, D. R., Rhee, S. G. & Williams, L. T. *Molec. cell Biol.* **10**, 2359-2366 (1990).
- Kaplan, D. R. et al. *Cell* **61**, 125-133 (1990).
- Margolis, B. et al. *Cell* **57**, 1101-1107 (1989).
- Turner, J. M. et al. *Cell* **60**, 755-765 (1990).

69. Barde, Y.-A., Edgar, D. & Thoenen, H. *EMBO J.* **1**, 549–553 (1982).
 70. Leibrock, J. *et al. Nature* **341**, 149–152 (1989).
 71. Hohn, A., Leibrock, J., Bailey, K. & Barde, Y.-A. *Nature* **334**, 339–341 (1990).
 72. Maisonnier, P. *et al. Science* **247**, 1446–1451 (1990).
 73. Alderson, R. F., Alterman, A. L., Barde, Y.-A. & Lindsay, R. M. *Neuron* **5**, 297–306 (1990).
 74. Rosenthal, A. *et al. Neuron* **4**, 767–773 (1990).
 75. Ernors, P. *et al. Proc. natn. Acad. Sci. U.S.A.* **87**, 5454–5458 (1990).
 76. Lindsay, R. M., Thoenen, H. & Barde, Y.-A. *Dev. Biol.* **112**, 319–328 (1985).
 77. Ernors, P., Wetmore, C., Olson, L. & Persson, H. *Neuron* **5**, 511–526 (1990).
 78. Maisonnier, P. *et al. Neuron* **5**, 501–509 (1990).
 79. Rodriguez-Tebar, A. & Barde, Y.-A. *J. Neurosci.* **8**, 3337–3342 (1988).
 80. Rodriguez-Tebar, A., Dechant, G. & Barde, Y.-A. *Neuron* **4**, 187–192 (1990).

81. Klein, R., Conway, D., Parada, L. & Barbacid, M. *Cell* **61**, 647–656 (1990).
 82. Klein, R., Martin-Zanca, D., Barbacid, M. & Parada, L. F. *Development* **109**, 845–850 (1990).
 83. Martin-Zanca, D., Barbacid, M. & Parada, L. F. *Genes Dev.* **4**, 683–694 (1990).

ACKNOWLEDGEMENTS. We thank J. Lown and M. Barkett for technical assistance, N. Patil for the pCMV5A construct, E. Johnson and J. Gornham for antibodies, M. Fortes and L. Van Houton for assistance with the figures, J. Brugge for thoughtful discussion, and members of the Chao laboratory for advice. This research was supported by the National Cancer Institute and the DHHS; the March of Dimes, the Andrew W. Mellon Foundation and the National Cancer Institute (B.L.H.); from the NIH, ACS, the Hirsch/Caulier Trust Fund, Parke Davis and the Dorothy Rodbell Foundation (M.V.C.).

LETTERS TO NATURE

Velocity structure of the ring nebula around supernova 1987A

Arlin P. S. Crotts* & Stephen R. Heathcote†

* Department of Astronomy, Columbia University, New York City, New York 10027, USA

† Cerro Tololo InterAmerican Observatory, Casilla 603, La Serena, Chile 1353

THE optical fading of supernova 1987A has revealed a parsec-sized nebula glowing in narrow atomic lines. The emission is presumed to come from circumstellar material, shed during the life of the progenitor star, illuminated by the ultraviolet flash from the explosion. This material is the closest to the supernova site that has been detected (apart from some transient radio emission immediately after the explosion¹) and therefore represents the youngest pre-supernova structure available for study. Here we report a series of high-resolution spectra of the nebula. The hollow, 1.7-arcsec-wide nebula has the velocity field of a ring expanding at 10.3 km s^{-1} , not that of a limb-brightened spheroid. Fainter nebulosity within 3 arcsec is expanding slightly faster, as indicated by redshifted emission from behind the supernova. The geometry of the emission can be used to infer structure, provided that light travel time and recombination delays are accounted for. Our analysis sets an upper limit to the velocity of the wind from the progenitor during its red supergiant (RSG) phase, and implies an RSG lifetime of $<4 \times 10^5 \text{ yr}$ and an interval of $\sim 2 \times 10^4 \text{ yr}$ between the end of the RSG phase and the supernova explosion.

Successive observations paint an increasingly detailed picture of the nebula. IUE detected circumstellar lines soon after the ultraviolet continuum decayed². Their evolution indicated a size for the emission line region of $\sim 2''$, optically confirmed by Wampler and Richichi³, who reported a velocity gradient across the nebula, but not its direction or structural details. Imaging showed the nebula elongated east-west in optical lines (but invisible in the continuum), and centred in a fainter reflection echo and emission-line nebulosity⁴. Finer images showed the central nebula to be fainter in its core⁵, probably a ring or limb-brightened spheroid. Hubble Space Telescope (HST) imaging showed this more clearly¹². For the central nebula's velocity structure, Wood and Faulkner⁶ reported a $2.5 \text{ km s}^{-1} \text{ arcsec}^{-1}$ gradient north-south and only $0.5 \text{ km s}^{-1} \text{ arcsec}^{-1}$ east-west. They did not describe the larger nebula, $5''$ across. These and our results differ significantly from a third study⁵.

We obtained spectra at resolution $R \approx 30,000$ and in good seeing (often $<1''$ full width half maximum (FWHM)) with the $f/8$ echelle spectrograph on the 4-metre telescope at Cerro Tololo InterAmerican Observatory (CTIO), on days 741–742, 759, 909–910, 1024–1025, 1122–1124 and 1403–1405 after core collapse. The data have the supernova continuum removed by simple interpolation across the narrow emission lines. All velocities are heliocentric. The following line groups were observed simultaneously and at most epochs: $\text{H}\alpha$, $[\text{N II}]$ lines at $6,548 \text{ \AA}$ and $6,584 \text{ \AA}$, and $[\text{S II}]$ at $6,717 \text{ \AA}$ and $6,731 \text{ \AA}$; $\text{H}\beta$

and $[\text{O III}]$ at $4,959 \text{ \AA}$ and $5,007 \text{ \AA}$; $[\text{O II}]$ at $3,726 \text{ \AA}$ and $3,729 \text{ \AA}$ and $[\text{Ne III}]$ at $3,869 \text{ \AA}$, and $[\text{N II}]$ at $5,754 \text{ \AA}$.

We first study the velocity field traced by the $[\text{N II}]$ $6,584 \text{ \AA}$ line on day 741 (Fig. 1). Its centroid velocity is 289.2 km s^{-1} . We find little velocity shear across the central nebula at position angle $\text{PA} = 60^\circ$, but large shears in opposing senses for position angles on either side, $\text{PA} = 20^\circ$ and 130° .

The morphology of the nebula arises from either a limb-brightened spheroid or an inclined and/or elongated ring. These have different implications for the nebula's kinematics. A circular ring with purely radial motion reproduces both the observed kinematics and morphology. To be consistent with HST imaging, such a ring would be inclined 43° from face-on with a semimajor axis, along $\text{PA} = 100^\circ$, of 245 light-days (assuming the distance of the Large Magellanic Cloud $D_{\text{LMC}} = 170,000$ light-years). An expansion velocity of 10.3 km s^{-1} reproduces the observed 14 km s^{-1} minor axis shear. The velocity field for this model (Fig. 1d–f), including seeing effects and spectrograph dispersion, matches the data at all position angles. The main deviations correspond to the larger echo loops discussed below.

This ring would be struck by the ultraviolet flash first on day 78, and last on day 409. These epochs correspond to when narrow ultraviolet lines 'turned on'², and later when many ultraviolet linestrengths began declining. This match, and the expansion velocity, would differ for an eccentric ring. This geometry also explains the lack of a reflection echo from the nebula on day 750 (ref. 4), despite more dimly fluorescent structures echoing strongly in the continuum. The nebula therefore extends less than ~ 300 light-days behind the supernova.

A sphere or prolate spheroid (with homologous velocity flow) generates no velocity shear on the projected minor axis, and oblate or triaxial solutions are highly constrained. To recover a shear of 14 km s^{-1} while limiting the velocity dispersion to $<3 \text{ km s}^{-1}$ as observed (with instrumental smearing deconvolved), an oblate or triaxial spheroid must be thinner than 1:5 in aspect ratio along $\text{PA} = 10^\circ$. For such thin oblate spheroids, the minor axis inclination approaches 43° , resembling the proposed ring. No such constraint exists for the triaxial case. These results are less sensitive than light-curve constraints to holes in the nebula's matter distribution.

In the same spectra, $\text{H}\alpha$ lines (Fig. 1g–i) show rough agreement with Fig. 1d–f, but there are details inconsistent with the model. $\text{H}\alpha$ being broader than $[\text{N II}]$ implies $T = 1.3 \times 10^4 \text{ K}$. Electron temperatures predicted for portions of the nebula on day 742 are $1\text{--}4 \times 10^4 \text{ K}$ (ref. 7), albeit for a spherical shell model. Cooling rates have yet to be calculated for a ring and depend on the H I recombination time (S.R.H. and A.P.S.C., manuscript in preparation). The ring appears hotter to the south, farther from Earth, as expected from light travel delay.

Apart from the ring we see a spatially and spectrally compact spot at $\text{PA} = 20^\circ$ (feature 'A' in Fig. 1g), corresponding to the $\text{H}\alpha$ source imaged on day 750 at $\text{PA} = 10^\circ$ (ref. 4). It vanished by day 909. Other spots are seen in images and spectra (also unresolved in velocity): feature 'B' in Fig. 1i (compare also ref. 8), and a fainter spot at $\text{PA} = 165^\circ$ in imaging on day 750 (ref. 4). All spots are just beyond the arclike echoes found within a

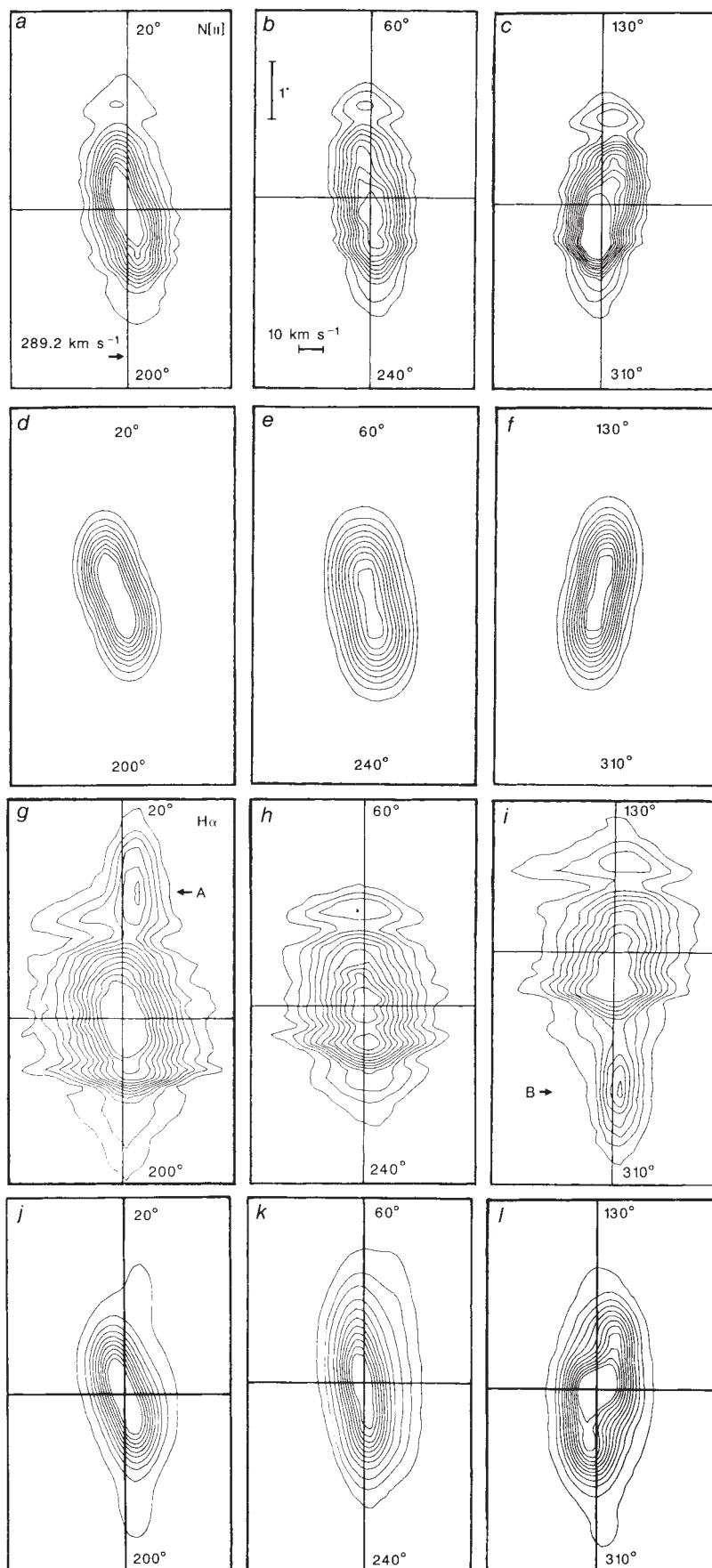


FIG. 1 Narrow [N II] emission at 6,584 Å around SN1987A on day 742 at different position angles of the spectrograph slit: *a*, 20°, *b*, 60° and *c*, 130°. The scale bars in *b* indicate units of slit length and velocity. Velocity increases to the right and the sense of the slit orientation is indicated by the arrow in *a*. The centroid velocity, 289.2 km s⁻¹, and the spatial position of the supernova are marked. *d*–*f*, As for *a*–*c*, but for the circular ring model described in the text. The same position angles are shown. *g*–*i*, The H α line on day 742, for the same position angles. The features labelled A and B are described in the text. *j*–*l*, [N II] line at 6,584 Å on day 1402, for the same position angles.

radius $r=3''$ and suggest instabilities along this interface between RSG and blue supergiant (BSG) winds. Elsewhere (S.R.H. and A.P.S.C., manuscript in preparation) we will use these spectra to estimate the properties of these clouds (as in ref. 4).

The faint, extended circumstellar features between these spots and the ring are important in understanding both. At radii $\sim 2''$ we see isolated line emission patches or extensions to the ring in both [N II] and H α . These correspond to arcs first imaged on day 750 (ref. 4) that evolved only slightly in position by $\sim$ day 910 (refs 5, 9).

During our observations, the ultraviolet flash propagated from $2.6''$ from the supernova at right angles to the line-of-sight to $4.8''$ by day 1403. If these structures glow because of prompt recombination after the ultraviolet flash, they fall near this right-angle plane. If the velocities (measured from their lines) are due to radial outflow from the progenitor, their components along the line of sight should be small. Small offsets from the centroid velocity are indeed measured for the features at 1.5 – $2.5''$. The tips of the ring in [N II] at PA = 20° , 130° , 200° and 310° stand $\sim 7 \text{ km s}^{-1}$ to either side of the centroid, but the fainter extensions beyond $\sim 2''$ do not continue the gradient established by the ring.

These offsets are expected to evolve as the ultraviolet flash sweeps backwards into the nebulosity. If the $2.5''$ radius nebula is well approximated by a sphere⁹ undergoing radial expansion, maximal change in redshift should have occurred during the interval of observation. Indeed, we see a shift of 7 km s^{-1} to the red during this time at PA = 20° (Fig. 1a and j).

The brightest arc is the only echo at $r < 2.5''$ and therefore the only echoing material far behind the supernova. In Fig. 1k, we see this feature separating from the ring, redshifted 13 km s^{-1} from the centroid. This material expands away from the supernova. From H α we find $T < 3,000 \text{ K}$ for this arc. The low temperature suggests that the velocity spread of the arcs at 1.5 – $2.5''$ ($\sim 25 \text{ km s}^{-1}$ FWHM) is due to different bulk velocities smeared together by recombination delays, not a thermal effect (with $T \approx 20,000 \text{ K}$). For instance, if the arcs expand at 15 km s^{-1} , then a $5 \times 10^7 \text{ s}$ recombination spread adds 12 km s^{-1} to the linewidth. This is supported by the subtle trend for red and blueshifted ends of the features in H α and [N II] to bend inward, with the largest radius at the centroid velocity. This would occur if the nebula were concave towards the supernova as in a sphere. Also, the fact that some of the features at $r \approx 2.5''$ are asymmetrical in velocity argues against a thermal effect.

These data show mass expanding from the star, so the central ring is probably expanding, with its north side towards Earth. Matter also expands outside the ring plane, perhaps into bipolar lobes, the south lobe expanding into the supernova foreground. This locus of matter is inconsistent with that allowed above for oblate or triaxial spheroids. Furthermore, the 'zig-zag' shape of isophotes in data obtained with good seeing, for example Fig. 1i, is reproduced by the ring model but not by a model of a thin spheroid.

Even a crude understanding of the expansion of these structures constrains the mass-loss history of the progenitor. If the central ring is due to collision between the RSG and BSG winds, the expansion velocity of the gas in the ring, 10.3 km s^{-1} , is an upper limit to the RSG flow velocity. We can then establish the time between the RSG-to-BSG transition and the supernova explosion by comparing the expansion velocity with the ring radius. The time obtained is 2×10^4 years; the value depends on poorly known details of the acceleration of ring material by the BSG wind.

The RSG wind terminates in a second shock, 12 light-years in radius. Applying the argument of ref. 10 to the minimum velocity of 10.3 km s^{-1} , we derive a relatively firm lower limit to the duration of the RSG phase of $4.5 \times 10^5 \text{ yr}$, with a mass loss rate of $2 \times 10^{-6} M_\odot \text{ yr}^{-1}$. There is about ten times as much mass accumulated in the terminus as there is still flowing

outward, so the actual RSG lifetime could be much longer⁹.

These time estimates agree with predictions from stellar evolution models^{10,11}. The RSG minimum lifetime is consistent with predicted values such as $8 \times 10^5 \text{ yr}$ (ref. 10), but uncomfortably close if ten times as much mass was lost before $t_{\text{sn}} - 4.5 \times 10^5 \text{ yr}$ as afterwards. Stellar evolution models must satisfy both the limits on the RSG lifetime and the course of the mass-loss rate of the progenitor. $\square$

Received 21 November 1990; accepted 6 March 1991.

1. Turtle, A. J. *et al.* *Nature* **327**, 38–40 (1987).
2. Fransson, C. *et al.* *Astrophys. J.* **336**, 429–441 (1989).
3. Wampler, E. J. & Richichi, A. *Astr. Astrophys.* **217**, L21–24 (1989).
4. Crotts, A. P. S., Kunkel, W. E. & McCarthy, P. J. *Astrophys. J.* **347**, L61–64 (1989).
5. Wampler, E. J., D'Odorico, S., Gouffes, C., Tarengi, M. & Wang, L.-F. *Astrophys. J.* **362**, L13–16 (1990).
6. Wood, P. R. & Faulkner, D. J. *IAU Circ. No.* 4739 (1989).
7. Lundqvist, P. & Fransson, C. *Astrophys. J.* (submitted).
8. Hanuschik, R. W. *Astr. Astrophys.* **237**, 12–16 (1991).
9. Crotts, A. P. S. & Kunkel, W. E. *Astrophys. J.* **366**, L73–77 (1990).
10. Woosley, S. E., Pinto, P. A. & Ensmann, L. *Astrophys. J.* **324**, 466–489 (1988).
11. Saio, H., Kato, M. & Nomoto, K. *Astrophys. J.* **331**, 388–393 (1988).
12. Jakobsen *et al.*, *Astrophys. J.* (in the press).

ACKNOWLEDGEMENTS. We thank E. Dwek and J. Felten for helpful discussions. This work began while A.P.S.C. was an NRC resident associate at Goddard SFC supported by NASA. Gerro Tolofo InterAmerican Observatory, NOAO, is operated by Associated Universities for Research in Astronomy, Inc. under contract with NSF.

The physical state of the intergalactic medium

X. Barcons*, A. C. Fabian† & M. J. Rees†

* Departamento de Física Moderna, Universidad de Cantabria, 39005 Santander, Spain

† Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

BECAUSE the process of galaxy formation is most unlikely to be perfectly efficient, there is a strong possibility that some baryonic gas remains outside collapsed structures such as galaxies and clusters of galaxies. What fraction of the baryonic content of the Universe resides in this intergalactic medium (IGM) and what physical state it is in are open questions. Here we use observational limits on the density of neutral hydrogen in the IGM, on the lack of deviations from a black-body spectrum of the cosmic microwave background (MBR), and on the extragalactic component of the soft X-ray background (XRB) to constrain the state of the IGM. From the lack of MBR fluctuations, any energetic IGM (containing as much energy as the binding energy in galaxies) is inferred to be smoothly distributed on scales greater than galactic. This rules out hot IGM models for the origin of the hard X-ray background, as well as the hypothesis that cosmic explosions may have given rise to cosmological structure on scales larger than galaxies.

When compared with the observed abundances of light elements, the standard theory of primordial nucleosynthesis (which assumes a perfectly homogeneous universe) places severe constraints on the fraction of the closure density of the Universe in the form of baryons¹: $0.04 \leq \Omega_{\text{baryon}} \leq 0.1$. Here we assume $H_0 = 50 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and $q_0 = 0.5$ throughout. The upper limit on Ω_{baryon} can be somewhat relaxed with the use of models invoking density inhomogeneities during the nucleosynthesis era². At redshift $z=0$, luminous matter in galaxies already contributes $\Omega_{\text{baryon}} \approx 0.007$ (ref. 3); a similar amount is in intra-cluster gas⁴, and some galactic or cluster baryonic dark matter could be present. A reasonable lower limit for the density of the IGM is $\Omega_{\text{IGM}} \geq 0.01$ at the present epoch.

There are, however, several recently updated observational facts that allow the density and temperature of the IGM to be constrained more directly, namely: (1) the lack of continuous absorption by neutral hydrogen (H I) shortward of the Lyman

α emission line in distant QSOs⁵ (the Gunn-Peterson effect); (2) the upper limits to the extragalactic XRB at 0.25 keV (ref. 6) and 1 keV (ref. 7); (3) the lack of distortion in the submillimetre spectrum of the MBR⁸; and (4) improved limits on small-angular-scale anisotropy of the MBR. In a previous, complementary study⁹, the properties of the ultraviolet ionizing flux were explored with the use of some of the above constraints. Here we use these constraints to gain information about the temperature of the IGM now and at an early epoch ($z=4.5$).

The existence of significant amounts of diffuse H I in the IGM was ruled out many years ago⁵ through measurements of the Gunn-Peterson effect. Diffuse H I should produce an absorption trough shortward of the Lyman α emission line in the spectra of QSOs, but this has not yet been detected (constraint (1)). This means that the IGM must be highly ionized out to the highest reachable redshifts ($z \approx 5$).

The most stringent limit on the Gunn-Peterson effect has been obtained recently from a quasar at $z \approx 2.64$ by taking into account discrete absorption by Lyman α forest lines¹⁰. The

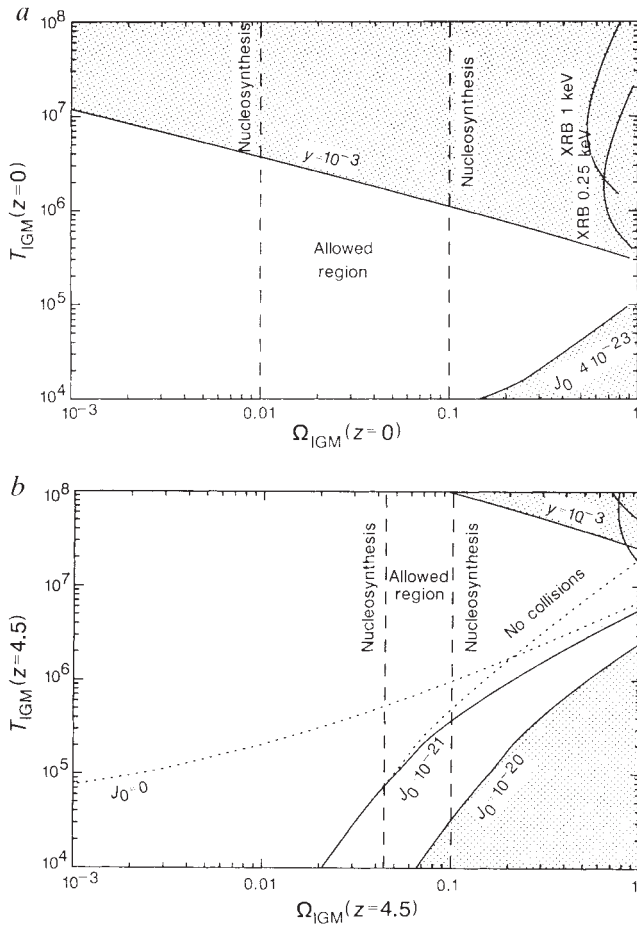


FIG. 1 *a*, Constraints on the temperature and density parameter for the IGM at the present epoch. The $y=10^{-3}$ line represents the upper limit to T_{IGM} from COBE, and the curves labelled XRB denote the maximum IGM density from current upper limits to the extragalactic X-ray flux. Both assume adiabatic cooling from $z=6$. The line marked $J_0=4 \times 10^{-23}$ is the minimum temperature as inferred from the absence of a Gunn-Peterson trough at $z=2.64$ with an ultraviolet flux $J_0(z=0)=4 \times 10^{-23} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$ evolving as $(1+z)^4$. The temperature of the IGM can be below the limit shown here only if there was a large energy input at the current epoch (so that the assumption of adiabatic cooling is incorrect). *b*, Same as *a* but at $z=4.5$. The IGM is assumed to be suddenly reheated at $z=4.5$ at the temperature shown. A Gunn-Peterson depression of less than 10% is assumed for several values of the ionizing flux at that redshift (in $\text{erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$). The minimum temperature is also shown for $J_0(z=4.5)=10^{-21} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$ and no collisional heating.

inferred upper limit on the density is $n_{\text{H I}} < 10^{-12} \text{ cm}^{-3}$ at that redshift. Two main processes can contribute to the ionization of the IGM: photo-ionization by ultraviolet sources and collisions. The ultraviolet flux is not known with any accuracy. At $z \approx 2-3$, QSOs produce a flux of only $J_0 \approx 10^{-21.7} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$ at the Lyman limit, but young galaxies (if not obscured by dust) could increase this number by a factor of ~ 10 (ref. 11). The reflection model for the XRB¹² predicts a value for J_0 of $\sim 2 \times 10^{-21} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$ at the same redshift. Estimates of the ionizing flux at the present epoch ($z=0$) from H I boundaries of galaxies¹³ suggest that $J_0 \approx 4 \times 10^{-23} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$. This increases as $\sim (1+z)^4$ back to the redshift at which the bulk of the background is generated; but if the predominant sources of ultraviolet radiation were at $z \geq 3$ then J_0 could even rise more steeply than $\sim (1+z)^4$ beyond $z \approx 3$, where Lyman-limit absorption affects all lines of sight¹¹.

In Fig. 1*a* we plot the minimum temperature required to keep the IGM ionized enough (that is, with $n_{\text{H I}} \leq 10^{-12} \text{ cm}^{-3}$) at $z=2.64$ as a result of collisions and photo-ionization, assuming a mean flux of $4 \times 10^{-23} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$ at $z=0$ which evolves as $(1+z)^4$. Ionization equilibrium is assumed¹⁴. The heating and cooling processes that we consider for determination of the $z=0$ temperature are adiabatic cooling, ultraviolet heating, collisional ionization and recombination cooling. As was discussed in ref. 9, this photo-ionization rate is sufficient to satisfy the Gunn-Peterson constraint unless $\Omega_{\text{IGM}} \geq 0.1$. If photo-ionization is still important at the present epoch, the temperature probably does not drop significantly below 10^4 K . Analogous limits on the IGM temperature are shown in Fig. 1*b* for $z=4.5$. We assume that the Gunn-Peterson trough there has a depth of less than 10%. The level of the ionizing background is highly uncertain at this very high redshift: there may be fewer photo-ionizing sources, but on the other hand absorption by intervening Lyman-limit clouds may mean that extrapolations of lower redshift estimates of J_0 are underestimated. If there were no heating input other than photo-ionization, T_{IGM} could not exceed $2 \times 10^4 \text{ K}$. Figure 1*b* therefore shows that with a modest ionizing ultraviolet flux ($J_0 \approx 10^{-21} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \text{ Hz}^{-1}$) Ω_{IGM} cannot exceed ~ 0.03 unless there is some collisional heating. This figure also shows explicitly how hot the gas must be for different values of J_0 . If J_0 were so small that photo-ionization were negligible, the temperature of the IGM at $z \sim 4.5$ would have to be at least a few times 10^5 K for Ω_{IGM} in the range 0.01–0.03. The lower limit to T rises roughly proportional to Ω_{IGM} for higher densities.

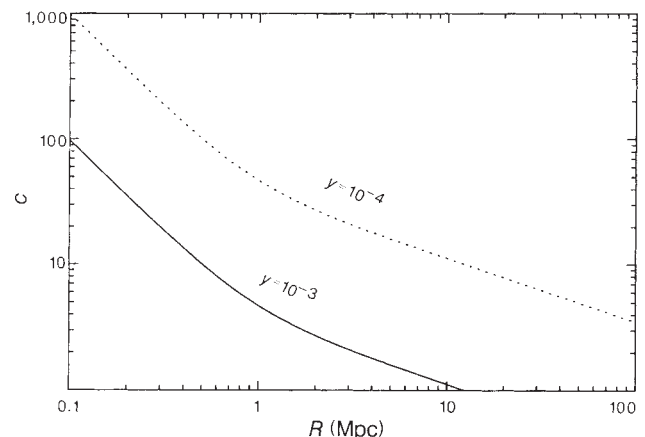
The most recent upper limits^{6,7} to the extragalactic soft X-ray intensity (constraint (2) above) are $I_X(0.25 \text{ keV}) \leq 60 \text{ keV cm}^{-2} \text{ s}^{-1} \text{ keV}^{-1}$ and $I_X(1 \text{ keV}) \leq 20 \text{ keV cm}^{-2} \text{ s}^{-1} \text{ keV}^{-1}$. The contribution of the IGM to the soft X-ray background depends on its clumpiness, characterized by the clumping factor

$$C = \left(\frac{\langle n_{\text{IGM}}^2 \rangle - \langle n_{\text{IGM}} \rangle^2}{\langle n_{\text{IGM}} \rangle^2} \right)^{1/2}$$

The bremsstrahlung volume emissivity is proportional to n_{IGM}^2 , so $I_X \propto (1+C^2)$. The temperature-dependent upper limits to Ω_{IGM} for a uniform IGM ($C=0$) are shown in Fig. 1, where the intensities have been calculated for a constant gaunt factor¹⁵, which provides a good approximation for our purposes. A reheating redshift $z_{\text{max}}=6$ (just enough to fulfill the Gunn-Peterson effect) and subsequent adiabatic cooling have been assumed for Fig. 1*a*.

Deviations from a purely planckian spectrum in the submillimetre region of the MBR are usually parameterized in terms of the comptonization y -parameter, which is defined as the average shift in the logarithm of the frequency for a microwave photon. The first results from the Cosmic Background Explorer (COBE)⁸ give an upper limit of $y < 10^{-3}$ (constraint (3)). The Compton distortion parameter, y , measures the integrated pressure along

FIG. 2 Maximum allowed values of the clumping factor C as a function of comoving clumping radius, R , for several values of the Compton distortion parameter, from $(\Delta T/T)_{\text{MBR}} < 2 \times 10^{-5}$ for a beamwidth of 1 arcmin (ref. 19).



the line of sight in the IGM, increased by a factor of $(1 + 4kT/m_e c^2)$, which is important at high temperatures. (We again assume simple adiabatic cooling for the gas, which allows us to place the upper limits shown in Fig. 1.)

These three observational constraints produce a picture of a uniform IGM consisting of a highly ionized gas with a temperature (at $z=0$) of $10^4 \text{ K} \leq T_{\text{IGM}} \leq 10^7 \text{ K}$. As a first conclusion, the hard X-ray background¹⁶ at 3–100 keV cannot arise as thermal bremsstrahlung in a hot uniform IGM, as that would require $T \geq 10^8 \text{ K}$ and $\Omega_{\text{IGM}} \approx 0.25$ (ref. 17), conflicting strongly with the COBE results.

At the more likely temperatures of $\sim 10^4 \text{ K}$, however, there are reasons to expect the IGM to be inhomogeneous. Strong localized sources (such as QSOs) will heat up their surroundings above the average, and deep, extended potential wells (such as clusters of galaxies) will concentrate high-density gas at temperatures of $\sim 10^7 \text{ K}$. Therefore, we must explore the possibility that the IGM is significantly clumpy.

To do this, we assume a simple model in which randomly distributed spherical blobs with gaussian gas-density profiles contain all of the IGM gas¹⁸. The evolution of the clumping factor C is uncertain, but is most important at the highest redshifts where most of the background is generated. We assume that C is independent of redshift.

An obvious consequence of clumping is that the soft X-ray background curves in Fig. 1 have to be shifted towards lower values of Ω_{IGM} by a factor of order $1/C$. Therefore, if $C \geq 10$ these constraints become relevant, and for $C \approx 100$ they require $\Omega_{\text{IGM}} \leq 0.01$. Furthermore, the soft X-ray background would also show important fluctuations on the clumping scale.

But a much stronger constraint than the isotropy of the soft X-ray background comes from the lack of angular fluctuations in the MBR (constraint (4)). If $C > 0$, the MBR will show fluctuations owing to the fact that microwave photons will be scattered by a finite number of IGM blobs along the line of sight. To our knowledge, the most stringent limit is $(\Delta T/T)_{\text{MBR}} \leq 2 \times 10^{-5}$ for a beamwidth of ~ 1 arcmin (ref. 19). The expected level of fluctuations in the MBR for a clumpy IGM is of the order of $(\Delta T/T)_{\text{MBR}} \sim y/\sqrt{N_b}$, where y is the total Compton distortion parameter and N_b is the total number

of IGM blobs per beam. Figure 2 shows the maximum clumping factor C on different scales coming from the above upper limit on $(\Delta T/T)_{\text{MBR}}$, where beamwidth effects have been properly included. The result is that significant clumping ($C \geq 10$) can occur only on scales $\leq (y/10^{-3})^{-1} \text{ Mpc}$.

This has a number of implications for specific models. First, there is the possibility already mentioned that the 3–100-keV X-ray background could arise as thermal bremsstrahlung in the IGM ($T_{\text{IGM}}(z=0) \geq 10^8 \text{ K}$). Even if the reheating redshift is as low as ~ 4 , the maximum amount of IGM at this temperature allowed by the COBE upper limit ($y < 10^{-3}$) is $\Omega_{\text{IGM}} \approx 0.01$, whereas $\Omega_{\text{IGM}} \approx 0.25$ is required for a uniform IGM that gives rise to the hard XRB^{17,18}. This results in a minimum clumping factor $C \geq 25$, and if upper limits on y decrease significantly, bigger clumping factors would be required. According to Fig. 2, this means that the hot gas should be confined to blobs with a size below $\sim 0.5 \text{ Mpc}$, which corresponds to the scale of a galaxy. No class of point-like source has yet been found that has a spectrum similar to the 3–100-keV X-ray background. Clearly, if the X-ray background arises from such a source, it is not from the IGM: the XRB must come from point sources.

A second important consequence concerns the explosion scenario for galaxy formation²⁰. In this picture, galaxies explode and throw an energy equivalent to their binding energy into the IGM. When the bubbles finally overlap, they must have typical sizes of several Mpc, because the model aims to reproduce the structure of the Universe up to $\sim 10 \text{ Mpc}$. Now, this model produces $y \approx 10^{-3}$ with very significant clumping on scales much greater than $\sim 1 \text{ Mpc}$, which is not allowed by the isotropy of the MBR.

A further possibility that has to be considered is that the metallicity of the IGM might be similar to that of intracluster gas (that is, $\sim 0.3 Z_{\odot}$, where $Z_{\odot}$ is the solar metal/hydrogen density ratio). This is expected if the ratio of galaxy to gas density in the Universe is constant. The possibility of detecting resonance absorption lines from highly ionized species (such as O VIII) in the IGM towards distant featureless X-ray sources (such as BL Lacs²¹) is then an important and observable diagnostic for the composition of the IGM. $\square$

Received 8 November 1990; accepted 28 January 1991.

1. Boesgaard, A. M. & Steingman, G. A. *Rev. Astr. Astrophys.* **23**, 319–378 (1985).
2. Malaney, R. A. & Fowler, W. A. *Astrophys. J.* **333**, 14–20 (1988).
3. Faber, S. M. & Gallagher, J. S. A. *Rev. Astr. Astrophys.* **17**, 135–187 (1979).
4. Sarazin, C. L. *Rev. mod. Phys.* **58**, 1–115 (1986).
5. Gunn, J. E. & Peterson, B. A. *Astrophys. J.* **142**, 1633–1636 (1965).
6. MacCammon, D. & Sanders, W. T. A. *Rev. Astr. Astrophys.* (in the press).
7. Wu, X., Hamilton, T., Helfand, D. J. & Wang, Q. *Astrophys. J.* (in the press).
8. Mather, J. C. et al. *Astrophys. J.* **354**, L37–L40 (1990).
9. Miralda-Escudé, J. & Ostriker, J. P. *Astrophys. J.* **350**, 1–22 (1990).
10. Steidel, S. & Sargent, W. L. W. *Astrophys. J.* **318**, L11–L14 (1987).
11. Bechtold, J., Weymann, R. J., Lin, Z. & Malkan, M. A. *Astrophys. J.* **315**, 180–197 (1987).

12. Fabian, A. C., George, I. M., Miyoshi, S. & Rees, M. J. *Mon. Not. R. astr. Soc.* **242**, 14P–16P (1990).
13. Kenney, J. D. P. in *The Interstellar Medium in Galaxies* (eds Thronson, H. A. & Shull, J. M.) 151–180 (Kluwer, Dordrecht, 1990).
14. Black, J. H. *Mon. Not. R. astr. Soc.* **197**, 553–563 (1981).
15. Sherman, R. D. *Astrophys. J.* **256**, 370–373 (1982).
16. Marshall, F. E. et al. *Astrophys. J.* **235**, 4–10 (1980).
17. Guilbert, P. W. & Fabian, A. C. *Mon. Not. R. astr. Soc.* **220**, 439–451 (1986).
18. Barcons, X. & Fabian, A. C. *Mon. Not. R. astr. Soc.* **230**, 189–206 (1988).
19. Partridge, R. B. in *Physical Cosmology* (ed. Trần Thanh Vân, J. (Editions Frontières, Gif-sur-Yvette, in the press).
20. Ostriker, J. P. & Cowie, L. L. *Astrophys. J.* **243**, L127–L132 (1981).
21. Canizares, C. R. & Kruper, J. *Astrophys. J.* **278**, L99–L102 (1984).

A nearby 37.9-ms radio pulsar in a relativistic binary system

A. Wolszczan

Arecibo Observatory, Arecibo, Puerto Rico 00613, USA

WITH very few exceptions¹, high Galactic latitudes have not been surveyed with sufficient sensitivity to detect millisecond pulsars. Accordingly we have conducted a preliminary sensitive survey at Galactic latitudes $b \geq 30^\circ$ using the Arecibo 305-m radiotelescope at a frequency of 430 MHz. Here I report the discovery of a 37.9-ms radio pulsar, PSR1534+12, in a 10.1-hour eccentric binary orbit². Timing analysis constrains the masses of the pulsar and its companion to be $1.32 \pm 0.03 M_\odot$ and $1.36 \pm 0.03 M_\odot$ (where $M_\odot$ is the mass of the Sun). This, together with the high eccentricity ($e = 0.274$) and small orbital diameter ($\sim 2 R_\odot$), indicates that the companion is probably another neutron star, and that the orbital evolution is strongly influenced by effects due to general relativity. The exceptionally high timing accuracy obtainable in PSR1534+12, because the pulse is so strong and narrow, will allow general relativity to be tested with unprecedented accuracy. The unique morphology of the pulsar's emission and polarization will make possible the measurement of a previously unconfirmed relativistic effect, the geodetic precession of the pulsar spin-axis.

The Arecibo survey was carried out in February 1990 during an extensive maintenance period, when the restrictions on the telescope motion limited its usage to transit observations. The ranges of Galactic longitudes $0^\circ \leq l \leq 35^\circ$ and $210^\circ \leq l \leq 360^\circ$ were drift-scanned for 4–5 h per day at constant elevation. This provided an integration time of 34 s within the half-power limits of the 10' telescope beam. The data were acquired in the manner described in earlier reports³. With the sampling interval of 506.625 μ s and the above integration time, this scheme generated 65,536 data samples per beam in each of the 128 frequency channels over a 10-MHz receiver bandwidth and provided a 6σ sensitivity of ~ 1.5 mJy down to pulsar periods of ~ 4 ms. In three weeks of observations, the survey covered an area of the sky ~ 150 square degrees. The data were analysed at the Cornell National Supercomputer Facility using the standard pulsar search procedure³.

The 37.9-ms pulsar (Fig. 1) was first detected in the data obtained on 4 February 1990. In addition, an analysis of the observations made on 9 February revealed the presence of another pulsar, PSR1257+12, with a period of 6.2 ms. Follow-up

TABLE 1 Parameters of the PSR1534+12 binary system

Pulsar period, P	$0.0379044403665 \pm 0.0000000000004$ s
Period derivative, $\dot{P}$	$2.43 \times 10^{-18} \pm 8 \times 10^{-20}$ s s ⁻¹
Right Ascension (B1950.0)	15 h 34 m 47.686 s ± 0.003
Declination (B1950.0)	$12^\circ 05' 45.23'' \pm 0.03$
Epoch	JD 2448200.0
Dispersion measure	11.61 ± 0.01 pc cm ⁻³
Flux density (430 MHz)	36 ± 8 mJy
Projected semi-major axis, $a_1 \sin i$	3.72943 ± 0.00003 light-s
Eccentricity, e	0.2736761 ± 0.0000007
Epoch of periastron, T_0	JD 2448199.8121643 ± 0.0000003
Orbital period, P_b	36351.7027 ± 0.0003 s
Longitude of periastron, ω	$264.672 \pm 0.0006^\circ$
Periastron advance, $\dot{\omega}$	$1.7562 \pm 0.0015^\circ$ yr ⁻¹
Time dilation, gravitational redshift, γ	0.0017 ± 0.0004 s
Orbital inclination, i	$\sim 74^\circ$
General relativity model masses and derived parameters	
Total mass, M	$2.679 \pm 0.003 M_\odot$
Companion mass, m_2	$1.36 \pm 0.03 M_\odot$
Pulsar mass, $m_1 = M - m_2$	$1.32 \pm 0.03 M_\odot$
Mass function $f(m_1, m_2)$	$0.3146 \pm 0.0002 M_\odot$
Pulsar surface magnetic field, B	1.0×10^{10} G
Pulsar characteristic age, τ_c	2.4×10^8 yr

timing measurements indicate that PSR1257+12 is probably a member of a long-period binary system. Here I will concentrate on the description of the PSR1534+12 system, postponing the discussion of PSR1257+12 until its parameters are known more accurately.

The shape and amplitude of the measured radial velocity curve shown in Fig. 2 indicate that PSR1534+12 is in an eccentric orbit around a massive companion. I obtained a phase-connected solution for the rotational and positional parameters of the pulsar, the keplerian orbit and the first two post-keplerian parameters $\dot{\omega}$ (rate of periastron advance) and γ (variable part of time dilation and gravitational redshift) from a least squares fit of the relativistic binary timing model of Damour and Deruelle^{4,5} to the observed arrival times of the pulses, using the analysis program TEMPO⁵ and the Center for Astrophysics Solar System ephemeris PEP740R. Pulsar parameters derived from the fit to the data spanning the period from February to December 1990 are listed in Table 1. The residuals after fitting of observations made with the bank of 32 contiguous 250-kHz filters and the Princeton Mark III pulsar processor at 430 MHz did not exceed 5 μ s for integration times of 2.5 minutes. Preliminary analysis indicates that reduction of this residual to 2 μ s or less is possible.

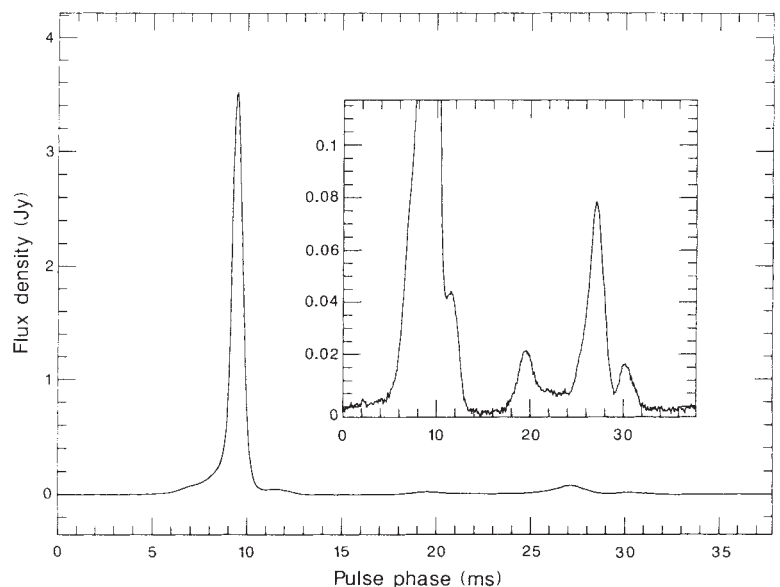


FIG. 1 The average pulse profile of PSR1534+12 at 430 MHz. The effective time resolution is ~ 300 μ s. The inset shows low-level emission enhanced by a factor of 20. The main pulse width at half-maximum is ~ 450 μ s.

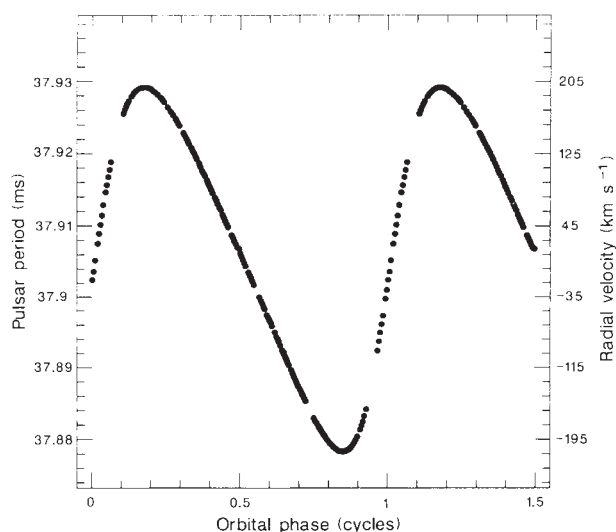


FIG. 2 The radial velocity curve of the orbital motion of PSR1534+12.

PSR1534+12 is evidently a member of a binary system whose parameters are similar to those of the canonical high-mass binary pulsar 1913+16 (ref. 5). From the facts that the minimum separation between the companion stars is $\sim 1 R_{\odot}$, that no eclipses are observed in the system and that the pulsar timing residuals show no significant deviations from the assumed point-mass binary model, it is almost certain⁶ that the companion of PSR1534+12 is another compact object. Moreover, the small slowdown rate of the pulsar implies that it is old and has a weak magnetic field (Table 1). These characteristics are typical of pulsars spun up by accretion of matter from binary companions and suggest that 1534+12 is the end-product of the evolution of a massive binary system⁷.

Masses of the two stars can be derived using an explicitly general-relativistic approach⁵ in which the post-keplerian orbital parameters are expressed in terms of a total system mass M and a companion mass m_2 . A fit of this model to the 1534+12 data gives accurate values of these parameters, $M = 2.679 \pm 0.003 M_{\odot}$ and $m_2 = 1.36 \pm 0.03 M_{\odot}$, immediately leading to a pulsar mass $m_1 = M - m_2 = 1.32 \pm 0.03 M_{\odot}$. These quantities are included in Table 1, and constraints imposed on the masses by the measured post-keplerian timing parameters in this model, and in the model in which $\dot{\omega}$ and γ are free parameters^{4,5}, are illustrated in Fig. 3.

The masses of PSR1534+12 and its companion are within the range of neutron star masses determined from observations of other radio and X-ray pulsars^{5,8-11}. They also agree with masses estimated from theoretical considerations of neutron star structure¹² and the mechanism of supernova explosions¹³. The component masses of the 1534+12 system agree well with the theoretically predicted 1.3–1.35 $M_{\odot}$ for the gravitational mass of a neutron star¹³. The masses in the 1913+16 system, 1.442 $M_{\odot}$ and 1.386 $M_{\odot}$ (ref. 5), and the 1534+12 data presented here are the most accurate determinations of neutron star mass so far.

As discussed above, the companion of PSR1534+12, apparently the heavier of the two objects, is very likely to be another neutron star. According to the standard model⁷, it was formed as the result of a collapse of its helium star progenitor. Assuming a spherical explosion, a straightforward calculation¹⁴ using the observed eccentricity and the pulsar mass leads to a progenitor mass of $\sim 2.1 M_{\odot}$. This is at the theoretical 'soft' lower limit to a helium core mass that would still collapse to a neutron star (2.2–2.5 $M_{\odot}$, refs 13, 15).

The observed eccentricity, as well as other orbital parameters, could be affected by an asymmetry in the supernova explosion of a more massive helium star^{14,16,17}. Preliminary modelling of the geometry of PSR1534+12 (A.W. and J. A. Phillips, manu-

script in preparation) using the observed polarization of the extended off-pulse emission (Fig. 1) indicates that the spin axis of the pulsar is inclined to the direction of the orbital angular momentum by $17^{\circ} \pm 2^{\circ}$ given the orbital inclination in Table 1. This result strongly suggests that the supernova explosion that gave rise to the pulsar's companion was indeed asymmetric.

In principle, the observed eccentricity could also result from a significant circularization of the orbit caused by gravitational radiation^{5,6} during orbit decay. Given the orbital parameters of PSR1534+12 and the masses of the companion stars, the theoretical rate of orbital decay ($\dot{P}_b \approx -1.9 \times 10^{-13}$) leads to a time for orbital dissipation ($\sim P_b/\dot{P}_b$) in excess of 10^9 yr. As this is much longer than the characteristic age of the pulsar ($\sim 2 \times 10^8$ yr), it is unlikely that the eccentricity observed now differs much from its value just after formation of the second neutron star.

PSR1534+12 exhibits a very rare combination of properties which make it a unique astrophysical object. Most notably, the pulsar's membership of a relativistic binary system offers new, fascinating opportunities to study gravity under strong-field conditions. The possibility of achieving timing accuracy of better than $2 \mu\text{s}$ for PSR1534+12, as compared with $\sim 20 \mu\text{s}$ for PSR1913+16 (ref. 5) and $\sim 65 \mu\text{s}$ for PSR2127+11C (ref. 9), means that relativity theories can be tested with unprecedented precision. Furthermore, the presence of a significant misalignment of the pulsar spin with the orbital angular momentum in the 1534+12 system means that the relativistic geodetic precession of the pulsar spin axis⁶ at the rate of $\sim 0.5^{\circ} \text{yr}^{-1}$ should manifest itself in the form of changes in the pulse shape, amplitude and polarization, as the pulsar beam drifts past the observer's line of sight. Results suggesting that this does occur in the 1913+16 system^{18,19} have not been conclusive in the absence of a constraint provided by the measurement of a spin-orbit angle. As this angle has already been measured for PSR1534+12, further long-term monitoring of its pulse arrival times, morphology and polarization should allow the existence of geodetic precession to be confirmed unambiguously.

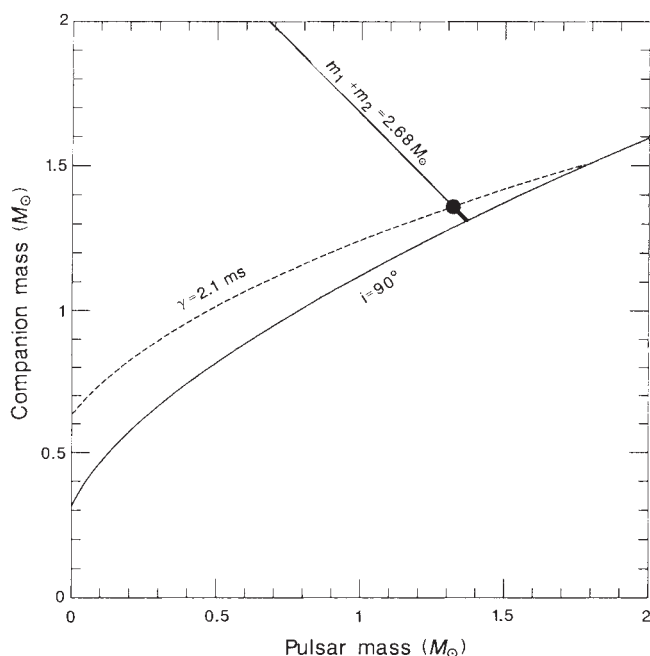


FIG. 3 General-relativistic constraints on masses of PSR1534+12 and its companion. A permissible range of masses (the thickened part of the $m_1 + m_2$ line) is delimited by the total mass derived from $\dot{\omega}$, the upper limit to γ and the $i = 90^{\circ}$ (edge-on orbit) condition imposed by the observed mass function. The masses obtained from the explicitly general-relativistic model are indicated by the filled circle, whose size reflects the estimation uncertainties.

Detections of PSR1534+12 and the 6.2-ms pulsar PSR1257+12 reported in this paper are important for our understanding of millisecond pulsar statistics^{20,21}. Both pulsars are at high Galactic latitudes, have low intrinsic luminosities (8–10 mJy kpc² at 400 MHz) and are at distances ≤ 0.5 kpc. Their discovery provides new evidence for the existence of a large population of old, 'spun up' pulsars in the Galaxy and indicates that future all-sky surveys have a good chance of detecting more millisecond pulsars in the solar neighbourhood. $\square$

Received 7 January; accepted 13 March 1991.

1. Stokes, G. H. *et al.* *Astrophys. J.* **311**, 694–700 (1986).
2. Wolszczan, A. *IAU Circular No.* 5073 (1990).
3. Wolszczan, A. *et al.* *Nature* **337**, 531–533 (1989).
4. Damour, T. & Deruelle, N. *Ann. Inst. H. Poincaré (Physique Théorique)* **44**, 263 (1986).
5. Taylor, J. H. & Weisberg, J. M. *Astrophys. J.* **345**, 434–450 (1989).
6. Smarr, L. L. & Blandford, R. *Astrophys. J.* **207**, 574–588 (1976).

7. van den Heuvel, E. P. J. & Taam, R. E. *Nature* **309**, 235–237 (1984).
8. Nagase, F. *Publ. astron. Soc. Japan* **41**, 1–79 (1989).
9. Anderson, S. B., Gorham, P. W., Kulkarni, S. R., Prince, T. A. & Wolszczan, A. *Nature* **346**, 42–44 (1990).
10. Lyne, A. G. & Bailes, M. *Mon. Not. R. astr. Soc.* **246**, 15P–17P (1990).
11. Ryba, M. F. & Taylor, J. H. *Astrophys. J.* **371**, in press.
12. Baym, G. & Pethick, C. J. *Ann. Rev. astr. Astrophys.* **17**, 415–443 (1979).
13. Woosley, S. E. in *The Origin and Evolution of Neutron Stars*, IAU Symp. 125 (eds Helfand, D. J. & Huang, J. H.) 255–272 (Reidel, Dordrecht, 1986).
14. Flannery, B. P. & van den Heuvel, E. P. J. *Astr. Astrophys.* **39**, 61–67 (1975).
15. Habets, G. M. H. J. *Astr. Astrophys.* **167**, 61–76 (1986).
16. Hills, J. G. *Astrophys. J.* **267**, 322–333 (1983).
17. Burrows, A. & Woosley, S. E. *Astrophys. J.* **308**, 680–684 (1986).
18. Weisberg, J. M., Romani, R. W. & Taylor, J. H. *Astrophys. J.* **347**, 1030–1033 (1989).
19. Cordes, J. M., Wasserman, I. & Blaskiewicz, M. *Astrophys. J.* **349**, 546–552 (1990).
20. Bhattacharya, D. & Srinivasan, G. *Curr. Sci.* **55**, 327–330 (1986).
21. Kulkarni, S. R. & Narayan, R. *Astrophys. J.* **335**, 755–769 (1988).

ACKNOWLEDGEMENTS. I thank J. H. Taylor for helpful comments and for making available the Mark III equipment and the latest version of TEMPO; also M. F. Ryba for assistance with the Mark III observing. Arecibo Observatory is part of the National Astronomy and Ionosphere Center, which is operated by Cornell University under contract with the NSF. Computations were performed at the Cornell National Supercomputer Facility which is supported in part by NSF, IBM Corporation, New York State and the Cornell Research Institute.

Anomalous neutron scattering near the superconducting phase transition

N. R. Bernhoeft*, P. J. Allen*, D. McK. Paul†, S. M. Hayden‡, P. A. Timmins‡ & G. G. Lonzarich§

* Physics Department, University of Durham, South Road, Durham DH1 3LE, UK

† Physics Department, University of Warwick, Coventry CV4 7AL, UK

‡ Institut Laue-Langevin, 156X, 38042 Grenoble Cedex, France

§ Interdisciplinary Centre on Superconductivity and Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK

THE inelastic scattering of low-energy neutrons near a second-order phase transition can provide a description of the spontaneous fluctuations in space and time of the physical parameter which, at low temperatures, enters a state of long-range order. Often the order parameter is coupled to the neutron directly, as in magnetic or lattice relaxation phenomena; as in the case of superconductivity, discussed here, the coupling is indirect. In either case, inelastic neutron diffraction measures the amplitude of the equilibrium fluctuations on an absolute scale, which may be compared with the predictions of models based on linear response theory¹. To this end, we report an investigation of the temperature dependence of small-angle neutron scattering in homogeneous bulk samples of the high-temperature superconductor YBa₂Cu₃O_{7-δ}. An anomalous increase in the scattering intensity occurs near the superconducting phase transition, and at lower temperatures. These effects seem to be larger than can be accounted for in elementary models of the normal and superconducting states.

The order parameter associated with the superconducting state may couple to the neutron in several indirect ways. Fluctuations in the amplitude and of the gradient of the phase of the order parameter generate slow fluctuations in the electron currents which couple to the neutron dipole moment. In a semi-classical picture, the currents act as sources of position- and time-dependent magnetic fields which diffract the neutron wave. If, on the appropriate space and time scales, the order parameter is also accompanied by fluctuations in the local spin density, this too may be seen as a source of fluctuating magnetic field². In addition, there are secondary sources of scattering in which the primary fluctuations of the order parameter modulate available degrees of freedom to which the neutron is directly sensitive. Thus, for example, longitudinal (charge-density) modulations of the superconducting order parameter may drive low-frequency ionic density oscillations which will be seen through the nuclear scattering potential, and modulations in the

order parameter will give rise to scattering from fluctuations induced in the normal electron fluid itself.

Superconducting materials with elevated transition temperatures offer a unique opportunity to gain insight into the dynamical properties of the superconducting order parameter by inelastic neutron scattering techniques. This arises for two reasons: first, the elevated temperatures allow a significant thermal population to be put into low-energy excitations; second, the contracted spatial scale of pairing, more or less implied by the high transition temperature (T_c) and the inferred magnitude of the effective binding energy allows measurements to be made at experimentally accessible scattering angles. We can thus explore fluctuations of low frequencies, $\omega \ll kT_c/\hbar$, and of characteristic wavelengths λ in the range $\xi_c < \lambda < \lambda_L$, where ξ_c is the superconducting pair coherence length and λ_L is the London screening length, without losing the response because of dynamical screening of magnetic fields (for wavelengths greater than λ_L) or the characteristic stiffening of the response (for wavelengths shorter than ξ_c).

Samples of YBa₂Cu₃O_{7-δ} were prepared as small, polycrystalline, sintered discs from materials of the highest quality, taking great care to maximize homogeneity at every stage of manufacture. Magnetization, resistive and X-ray studies confirmed their purity, phase composition and superconducting nature at temperatures below 91 ± 1 K.

The temperature dependence of small-angle scattering cross-section from the high-purity YBa₂Cu₃O_{7-δ} samples is shown in Fig. 1. The data show a clear increase in intensity over the background level as the material passes through the superconducting transition. This scattering is broadly distributed about the transition temperature over a range of ~ 10 K, which is more than an order of magnitude wider than the bulk screening transition region measured either inductively or resistively. The intensity of the excess scattering near T_c decreases with increasing wavevector. Below the phase transition a broad anomaly with a peak at ~ 30 K is seen. This too decreases in intensity with increasing wavevector.

The interpretation of the data fall naturally into two parts: first, that at elevated temperatures where the increased scattering signals the build-up of short-range order and the transition to the condensed state, and second, the scattering observed in the state of long-range order below T_c . We focus primarily on the changes taking place close to or just above the transition temperature, where long-wavelength and low-frequency fluctuations of the order parameter are normally most pronounced.

For reference, we first examine elementary models for the scattering predicted from spin-density (i) and current-density (ii) fluctuations for a normal Fermi liquid. We then consider fluctuations of the currents (iii) and of the lattice (iv) associated with the build up of short-range order. The results rule out the

simplest treatments of these mechanisms for the observed scattering and point to possible ways forward.

In models (i), (ii) and (iii) the focus of attention is the coupling of the neutron magnetic moment with the local dynamical magnetic field $\mathbf{B}(\mathbf{r}, t)$ generated by fluctuations in the electron spin and current densities. The connection between the scattering cross-section and the power spectrum of the field fluctuations will be considered for an unpolarized beam and an isotropic body in a Schrödinger–Pauli model¹. In this case the probability, per atom, incident flux, solid angle Ω and frequency ω , of neutrons being scattered from initial momentum $\hbar\mathbf{k}$ to final momentum $\hbar\mathbf{k}' = \hbar(\mathbf{k} + \mathbf{q})$ with energy increase $\hbar\omega$ can be expressed as

$$\frac{d^2\sigma}{d\Omega d\omega} = \frac{\gamma^2 e^2 k'}{48\pi^3 \hbar^2 c^2 n k} \langle |\mathbf{B}(\mathbf{q}, \omega)|^2 \rangle \quad (1)$$

where γ is the neutron gyromagnetic factor (1.913), e is the electronic charge, c is the speed of light, n is the number of atoms per unit volume, $k = |\mathbf{k}|$, $k' = |\mathbf{k}'|$ and $\mathbf{B}(\mathbf{q}, \omega)$ is defined as $v^{-1/2}$ times the Fourier transform $\int d^3r \int dt \mathbf{B}(\mathbf{r}, t) \exp[-i(\mathbf{q} \cdot \mathbf{r} - \omega t)]$ of $\mathbf{B}(\mathbf{r}, t)$ defined over v , the product of a large volume V and a long time interval. The quantity $\langle |\mathbf{B}(\mathbf{q}, \omega)|^2 \rangle = \langle \mathbf{B}(-\mathbf{q}, -\omega) \cdot \mathbf{B}(\mathbf{q}, \omega) \rangle$ is the power spectrum equal to the Fourier transform of the unsymmetrized autocorrelation function $V^{-1} \int d^3r' \langle \mathbf{B}(\mathbf{r}', 0) \cdot \mathbf{B}(\mathbf{r}' + \mathbf{r}, t) \rangle$.

The cross-section $d\sigma/d\Omega$ for total scattering in a given direction is obtained by integrating $d^2\sigma/d\Omega d\omega$ over ω at fixed scattering angle with $\mathbf{q}$ (and ω) satisfying the conservation condition $\hbar\omega + \hbar^2 k^2/2M = \hbar^2 k'^2/2M$, $\hbar\mathbf{k}' - \hbar\mathbf{k} = \hbar\mathbf{q}$, where M is the neutron mass. The quantity plotted in Fig. 1 is the change of $d\sigma/d\Omega$ relative to a q -dependent background of ~ 10 barn steradian⁻¹ atom⁻¹, averaged over the azimuthal angle of scattering. The use of low-energy neutrons and small scattering angles near the forward direction ensures that only slow fluctuations of long wavelength, which are those of special interest near T_c , contribute to the measured cross-section.

We may express the power spectrum of field fluctuations as the sum of two terms:

$$\langle |\mathbf{B}(\mathbf{q}, \omega)|^2 \rangle = (4\pi)^2 \frac{6\hbar\omega}{e^{\hbar\omega/k_B T} - 1} \cdot \left[\frac{\text{Im } \chi}{\omega} + K \frac{\text{Re } \sigma}{c^2 q^2} \right] \quad (2)$$

The first is the familiar ‘magnetic’ or spin-scattering term which is most conveniently expressed in terms of the absorptive (imaginary) part of the magnetic spin susceptibility $\chi(\mathbf{q}, \omega)$. The second is the orbital or current fluctuation term which is proportional to the real (absorptive) part of the transverse conductivity. The factor $K = K(\mathbf{q}, \omega)$ is the dynamic screening factor, which is of decisive importance at very long wavelengths and low frequencies and is given by

$$K(\mathbf{q}, \omega) = \frac{\mu^2}{[1 - \mu\omega^2/c^2 q^2 + 4\pi\omega\mu \text{Im } \sigma/c^2 q^2]^2 + [4\pi\omega\mu \text{Re } \sigma/c^2 q^2]^2} \quad (3)$$

where $\mu = \mu(\mathbf{q})$ is the wavevector-dependent static magnetic permeability, and $\sigma = \sigma(\mathbf{q}, \omega)$ is the transverse conductivity with real and imaginary parts, $\text{Re } \sigma(\mathbf{q}, \omega) + i \text{Im } \sigma(\mathbf{q}, \omega)$ respectively. The two contributions (spin and current fluctuation) to the magnetic-field power spectrum can be readily estimated for the normal Fermi liquid state^{3–8}. A full numerical analysis indicates that both mechanisms are at least two orders of magnitude too weak to account for the observed scattering near T_c .

As T_c is approached the Fermi liquid description is naturally expected to break down, and we now consider the extent to which the conventional description of this breakdown helps us to understand the enhanced scattering observed. The formation of virtual Cooper pair states, or more precisely the growth of short-range superconducting order, leads to a local enhancement

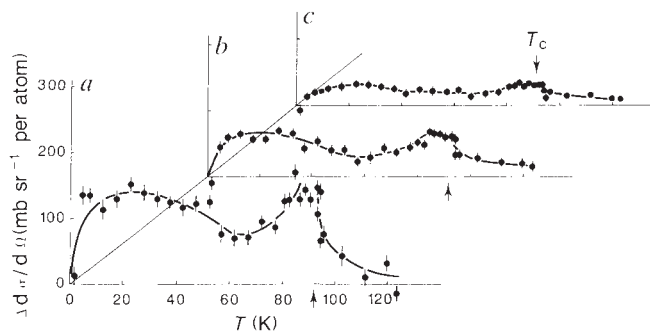


FIG. 1 Differential neutron scattering cross-section of high-homogeneity and high-purity polycrystalline samples of $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ against temperature, in absolute units of millibarns per steradian per average atom. The measurements were made on the small-angle spectrometer D11 at the Institut Laue–Langevin, with incident wavelengths of 6.5 Å. In the three frames *a*, *b* and *c* the angles of scattering correspond to elastic scattering wave vectors of 0.035 Å⁻¹, 0.045 Å⁻¹ and 0.055 Å⁻¹, respectively. The cross-sections are averages over the azimuthal angle of scattering and are given relative to a q -dependent background level of ~ 10 barn steradian⁻¹ atom⁻¹. Angular normalization and absolute calibration of the data has been carried out with the use of standard plexiglass and vanadium samples, respectively. The arrows indicate the value of T_c obtained from measurements of the bulk diamagnetic susceptibility.

of the conductivity at low frequencies and hence to a build up of slow fluctuations of the current density (case (iii)). The local conductivity is increased not only by the enhanced mobility of the virtual pairs, but also by an increased mobility of the unpaired carriers, whose motion near regions of emerging order is modified from that of the normal state. Current fluctuations arising from the enhanced mobility of virtual pairs in an isotropic state may be determined directly in terms of fluctuations of the complex order parameter $\psi(\mathbf{r}, t) = |\psi| \exp(i\theta)$, which yields the local pair concentration from $|\psi|^2$ and the local supercurrent density $\mathbf{J}_s(\mathbf{r}, t)$ for pairs of charge-to-mass ratio e/m through the relation $\mathbf{J}_s = \hbar e/2im(\psi^* \nabla \psi - \psi \nabla \psi^*)$. When the gradient of $|\psi|$ can be ignored, $\mathbf{J}_s$ reduces to $(\hbar e/m)|\psi|^2 \nabla \theta$ and is therefore proportional to the local concentration of supercarriers and the local gradient of the phase $\theta(\mathbf{r}, t)$ of the order parameter. The current power spectrum $\langle |\mathbf{J}_s(\mathbf{q}, \omega)|^2 \rangle$, and the corresponding conductivity and neutron scattering cross-section, can be directly evaluated from the statistical properties of $\psi(\mathbf{r}, t)$ normally described in terms of a time-dependent Ginzburg–Landau model. In the conventional analysis the fluctuations of $\psi(\mathbf{q}, t)$ are represented in terms of a gaussian stochastic process characterized by a single relaxation rate $\Gamma(\mathbf{q})$, expanded up to second order in $\mathbf{q}$, which softens at low $\mathbf{q}$ as T_c is approached^{9,10}.

Using this framework, a standard calculation yields a contribution to the conductivity that has qualitatively the correct temperature dependence (inverse power $(T - T_c)$). But if the parameters of the time-dependent Ginzburg–Landau model are taken to be those of the Bardeen–Cooper–Schrieffer theory for the electron gas, the cross-section is, like that of the normal state, far too weak to account for the observed scattering. This conclusion is not substantially changed when the enhanced mobility of the unpaired carriers is taken into account within existing frameworks^{11–13}.

So far we have considered only the coupling of the magnetic moment of the neutron to the fluctuating magnetic fields. The neutron also interacts directly with the nuclei (case (iv)). Because of the low incident energies and small scattering angles used, calculations based on the conventional analysis of lattice vibrations (single or multiple scattering involving zero, one or more phonons) cannot account for the observed intensity. Furthermore the observed fall in cross-section with increasing wavevector and temperature above T_c is not reproduced in such models.

The above considerations do not definitely rule out any of the three general scattering mechanisms (spin, current or nuclear). They merely indicate that the simplest forms of these mechanisms, within the limitations of the isotropic model considered, fail to give satisfactory quantitative agreement with the measured cross-sections. In the case of spin scattering, we may question the assumption of negligible susceptibility $\chi(\mathbf{q}, \omega)$: particularly for wavelengths shorter than the superconducting coherence length but longer than the spin correlation length, an enhancement of the susceptibility near T_c could affect both the spin and current scattering cross-sections and thus affect cases (i), (ii) and (iii). The second point of weakness arises in the usual analysis of supercurrent fluctuations, in which a phenomenological model is treated in a gaussian approximation with the parameters based on a mean-field theory and the electron-gas model, and in which the effects of interactions of the unpaired carriers with the fluctuations of the order parameter are incompletely developed. The indirect effects that short-range

order near T_c has on the displacements of the atoms from their normal equilibrium position (case (iv)) also need further investigation; a complete analysis appropriate for the conditions of our experiments has not yet been carried out. Finally, there exist interpretations of the data, not considered here, that rely on detailed microscopic modelling, for example quasielastic scattering from fields of persistent currents due to vortex lines or superconducting clusters and, in particular, scattering due to fluctuations in the angular momentum of pairs in anisotropic states.

The anomalous scattering we have observed appears to be associated with the dynamics of the superconducting order parameter, but standard models still cannot give an elementary account of the magnitude of the scattering. At this stage it is unclear whether the effect is unique to high- T_c superconductivity, with its quasi-two-dimensional character and possibly exotic quasiparticles. To this end we have initiated a programme of study of more conventional superconductors. $\square$

Received 18 January; accepted 15 March 1991.

1. Marshall, W. & Lovesey, S. W. *Theory of Thermal Neutron Scattering* (Oxford University Press, 1971).
2. Joyn, R. & Rice, T. M. *Phys. Rev.* **B38**, 2345–2353 (1988).
3. Bernheft, N. R., Hayden, S. M., Lonzarich, G. G., Paul, D. McK. & Lindley, E. J. *Phys. Rev. Lett.* **62**, 657–660 (1989).
4. Bezinge, A., Jorda, J. L., Junod, A. & Muller, J. *Solid St. Commun.* **64**, 80–82 (1987).
5. Landolt-Börnstein *Magnetic Properties* vol. 2 (Springer, Berlin, 1962).
6. Martin, P. C. *Phys. Rev.* **161**, 143–155 (1967).
7. Sasaki, K. & Obata, Y. *Suppl. Prog. theor. Phys.* **69**, 406–419 (1980).

8. Lindhard, J. *Mat.-Fys. Meddr Kongelige danske Videnskabernes Selskabs* **28.8** (1954).
9. Aslamasov, L. G. & Larkin, A. I. *Phys. Lett.* **A26**, 238 (1968).
10. Schmidt, H. Z. *Phys.* **216**, 336–345 (1968).
11. Maki, K. *Prog. theor. Phys.* **39**, 897–906 (1968).
12. Thompson, R. S. *Phys. Rev.* **B1**, 327–333 (1970).
13. Maki, K. & Thompson, R. S. *Phys. Rev.* **B39**, 2767–2770 (1989).

ACKNOWLEDGEMENTS. We thank the staff at the Institut Laue-Langevin for their cooperation and a number of our colleagues, in particular Y. Chen, P. Coleman, U. Cox, M. Delap, M. Gunn, A. Huxley, H. Mook, G. Squires, W. Stirling and J. Waldram, for their interest and for helpful discussions. This study was supported by the SERC.

Low organic carbon accumulation rates in Black Sea sediments

S. E. Calvert*, R. E. Karlin†, L. J. Toolin‡,
D. J. Donahue‡, J. R. Southon§¶ & J. S. Vogel§¶

* Department of Oceanography, University of British Columbia, Vancouver V6T 1W5, Canada

† Mackay School of Mines, University of Nevada, Reno, Nevada 89557, USA

‡ NSF-Arizona AMS Facility, University of Arizona, Tucson, Arizona 85721, USA

§ Department of Archaeology, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

THE Black Sea, the world's largest anoxic marine basin, is frequently used as a modern analogue for the formation of organic-rich sediments and carbonaceous rocks^{1–3}, on the widely held assumption that anoxic conditions promote the preferential preservation of organic matter in sediments. Data for testing this hypothesis have so far been equivocal^{4–7}, but here we use radiocarbon ages obtained using accelerator mass spectrometry for the organic fraction of recent Black Sea sediments to estimate the organic carbon accumulation rates. These range from 0.69 to 2.09 g C m⁻² yr⁻¹ and are significantly lower than earlier estimates based on varve counting⁶. Depending on the value taken for the rate of primary production in the Black Sea^{4,8}, between 0.7 and 2.1% of the organic carbon is preserved in the bottom sediments. When compared with carbon accumulation rates in equivalent oxygenated environments⁹, these results indicate that the modern Black Sea is not a site of anomalously high organic carbon accumulation. This suggests that anoxic conditions in the water column may not be a prerequisite for the preservation of organic matter in marine sediments, and that models of the origin of carbonaceous facies in the geological record may therefore need to be modified.

The Holocene section in the Black Sea¹⁰ comprises a modern, finely laminated coccolith marl (unit 1) overlying a microlaminated organic-rich sapropel (unit 2) which, in turn, overlies moderately laminated calcareous clays (unit 3). The units are thought to reflect changes in sedimentation first from a Pleistocene freshwater lake (unit 3) to an early phase of the present anoxic, brackish basin (unit 2) formed when the Black Sea was reconnected with the Mediterranean by the Bosphorus Strait after the post-glacial sea-level incursion, then to the present highly anoxic marine basin (unit 1).

The cores were collected during Leg 4 of the cruise of RV *Knorr* of the Woods Hole Oceanographic Institution in June–July 1988 (Fig. 1). We used a Kastengreifer¹¹ box-corer to collect undisturbed samples of the modern unit 1 sediments. Smaller sub-samples of the box cores were taken by inserting 10-cm-diameter butyrate core tubes into the boxes. Before removing the sub-cores from the boxes, we collected samples of the superficial 'fluff' layer¹² by carefully syringing off the upper few millimetres of the flocculent sediment from the undisturbed surfaces of some of the box cores. The sub-cores were processed by extruding and centrifuging successive 1- or 2-cm sections and then freeze-drying the solids. At each level, we also collected small sub-samples to determine water content.

In the shore laboratory, we carried out carbon analysis and radiocarbon dating on the sub-samples. The total and carbonate carbon values were determined by combustion-gas chromatography and by coulometry, with precisions of ± 1.2 and 3.7% (2 σ), respectively. Organic carbon was taken to be the difference between total and carbonate carbon. We calculated calcium carbonate contents from the carbonate carbon values assuming no other carbonate-bearing phase was present. Samples for radiocarbon dating were treated overnight with 10% HCl, washed three times with distilled water, dried and pulverized. The methods used for the radiocarbon assays of cores BS4-9BC and BS4-14BC at Arizona are given in refs 13 and 14, and those for core BS4-15BC at McMaster University in refs 15 and 16.

Figure 2 shows that the sediments recovered from unit 1 in the abyssal area of the Black Sea (cores BS4-9BC and BS4-14BC)

¶ Present address: Lawrence Livermore National Laboratory, Livermore, California 94550, USA.

TABLE 1 Sedimentation rates, sediment properties and organic carbon accumulation rates

Core number	Water depth (m)	Sedimentation rate (cm per 1,000 yr)	Grain density (g cm^{-3})*	Porosity†	Total sediment accumulation rate ($\text{g m}^{-2} \text{yr}^{-1}$)	Organic carbon (%)‡	Organic carbon accumulation rate ($\text{g C m}^{-2} \text{yr}^{-1}$)
BS4-9BC	2,087	15.9	2.42	0.91	38.7	6.04	2.16
BS4-14BC	2,218	15.8	2.43	0.91	35.7	5.07	1.81
BS4-15BC	198	17.0	2.48	0.90	42.2	1.64	0.69

* Calculated assuming the following grain densities: CaCO_3 2.7 g cm^{-3} ; organic matter (organic carbon $\times 2$) 1.0 g cm^{-3} ; clay ($1 - (\text{CaCO}_3 + \text{organic matter})$) 2.5 g cm^{-3} .

† Obtained from the water contents, corrected for the presence of salt in the dried samples. Mean values for the intervals below the marked decrease in organic C content (Fig. 1).

‡ Mean values for the intervals below the marked decrease in organic C content (Fig. 1).

have organic carbon contents that decrease from surface values of 9.6% to 4–7% at depths of 1 cm. Core BS4-15BC, recovered from a water depth of 198 m, has lower organic carbon contents. The radiocarbon ages of the organic fraction show that the surface 'fluff' samples are either essentially modern or contain bomb radiocarbon. Below the surface layers, the radiocarbon ages increase to $\sim 3,000$ years BP (before present) at the bases of the box cores. Evidence from gravity cores collected at the same sites indicates that the bottoms of the box cores were very close to the boundary between units 1 and 2. Regression plots of age against depth for the three cores (Fig. 2) have surface intercepts of 593–1,215 years BP, and slopes that yield sedimentation rates of 15.8–17.0 cm per 10^3 yr (Table 1).

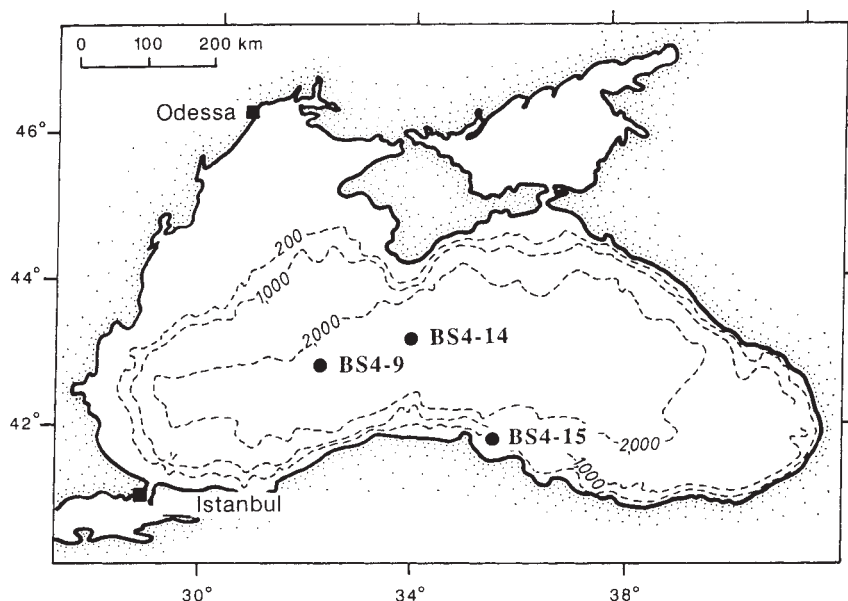
The age of the unit 1/unit 2 boundary in the Black Sea has been estimated previously at 3,000 BP from conventional radiocarbon dating of bulk organic matter¹⁰, and at 1,000 BP from varve counting¹⁷. The latter value was based on the assumptions that the fine-scale laminations in unit 1 are strictly annual deposits (varves)^{18,19} and that the complete section is preserved. Degens *et al.*¹⁷ argued that the radiocarbon ages are affected by old, presumably reworked, organic and/or carbonate carbon in the sediments, and that 2,000 years should be subtracted from ^{14}C ages to make them conform with the varve ages: in the absence of surface samples, this argument was accepted⁵. The new data for unit 1 (Fig. 2) show that the extrapolations of the age/depth curves to the surfaces of the cores yield ages that are substantially less than 2,000 years; a correction for the apparent radiocarbon age of the surface waters, probably lying between the value of 400 years adopted for the open ocean²⁰ and the

value of 1,430 years estimated from the ^{14}C balance in the Black Sea²¹, would lower these ages further. Hence, any contribution of old reworked carbon in these sediments is probably small; although some of the organic fraction of the unit 1 sediments may be derived from terrestrial sources⁵, this material is not necessarily older than the sediments into which it is mixed. In addition, radiocarbon ages of coccolith carbonate from unit 1 sediments, derived by accelerator mass spectrometry^{5,22} are similar to those of the organic carbon fraction. It seems unlikely that the admixture of old, reworked organic material and carbonate into the sediments would produce such similar radiocarbon ages for these fractions. We therefore interpret the agreement between carbonate and organic carbon ages as further evidence that there is very little old reworked organic or carbonate carbon in the unit 1 sediments, thus questioning the accuracy of the varve chronology.

Given the relative youth of the core surfaces as extrapolated from ^{14}C measurements, we assumed that the differences between the true and radiocarbon ages were sufficiently small and constant over the dated core intervals that the sedimentation rates derived from Fig. 2 were correct, and we used the radiocarbon dating to derive total sedimentation and organic carbon accumulation rates (Table 1). In these calculations, the organic carbon contents from below the organic-rich surface layer were adopted as the values most relevant to long-term burial. The total mass accumulation rates are $34.5\text{--}42.2 \text{ g m}^{-2} \text{yr}^{-1}$ and the organic carbon accumulation rates range from $0.69 \text{ g C m}^{-2} \text{yr}^{-1}$ at 198 m to $1.75\text{--}2.09 \text{ g C m}^{-2} \text{yr}^{-1}$ at 2,000 m.

Measurement by ^{210}Pb determination²³ of the sedimentation

FIG. 1 Locations of the box cores used in this study. Isobaths are in metres.



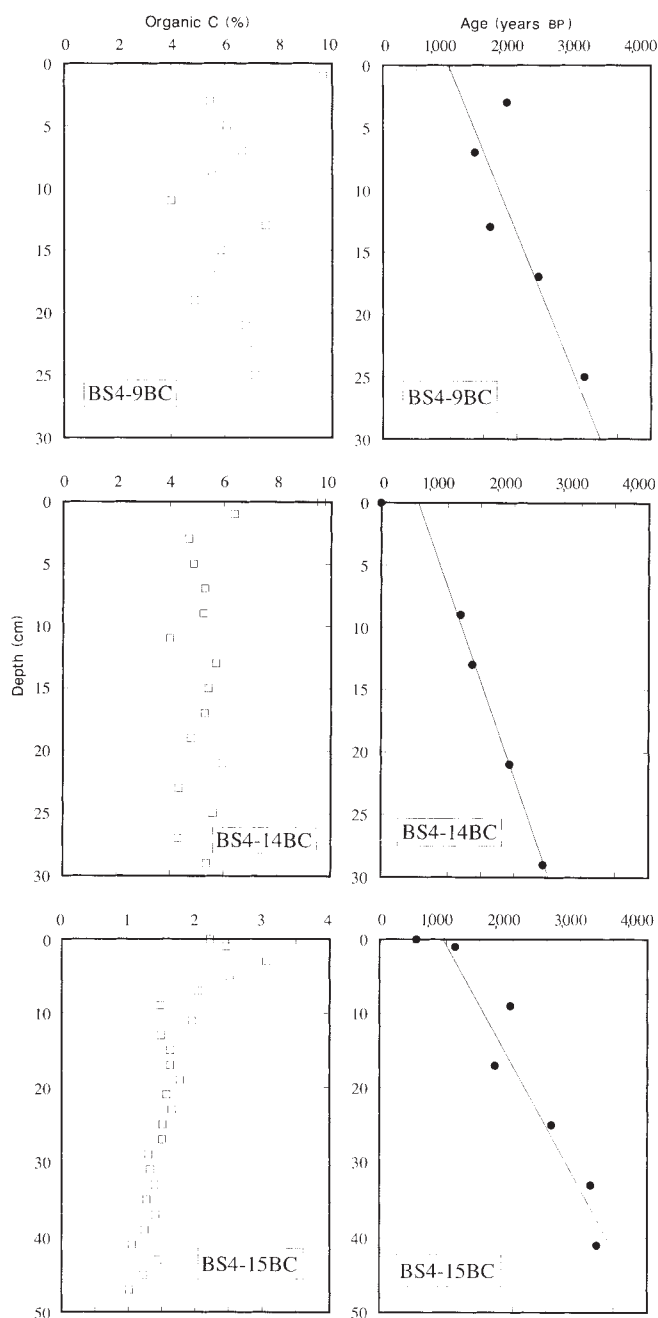


FIG. 2 Depth distributions of organic carbon and radiocarbon ages in three box cores from the Black Sea. Samples at a depth of 0 cm represent the 'fluff' layer. The surface 'fluff' sample in core BS4-14BC has been omitted from the regression for this core because it contains bomb ^{14}C .

rate in the upper 3 cm of core BS4-9 yielded a total sediment accumulation rate of $55 \text{ g m}^{-2} \text{ yr}^{-1}$ which is compatible with the $34.6 \text{ g m}^{-2} \text{ yr}^{-1}$ determined by the radiocarbon method over a longer time interval. Our calculated organic carbon accumulation rates are similar to those of between 0.1 and $4.8 \text{ g C m}^{-2} \text{ yr}^{-1}$ reported²⁴ for water depths below 1,000 m. Furthermore, total mass accumulation rates of unit 1 sediments based on varve counting⁶ range from 150 to $171 \text{ g m}^{-2} \text{ yr}^{-1}$, and yield organic carbon accumulation rates of $6.0\text{--}7.0 \text{ g C m}^{-2} \text{ yr}^{-1}$. These are higher than the settling fluxes of organic carbon ($1.4\text{--}1.8 \text{ g C m}^{-2} \text{ yr}^{-1}$) measured by moored sediment traps at two stations at 1,200-m depth in the Black Sea¹⁹. These data, together with the detailed comparison²³ between lamina counting and ^{210}Pb determinations, further indicate that the assumption of

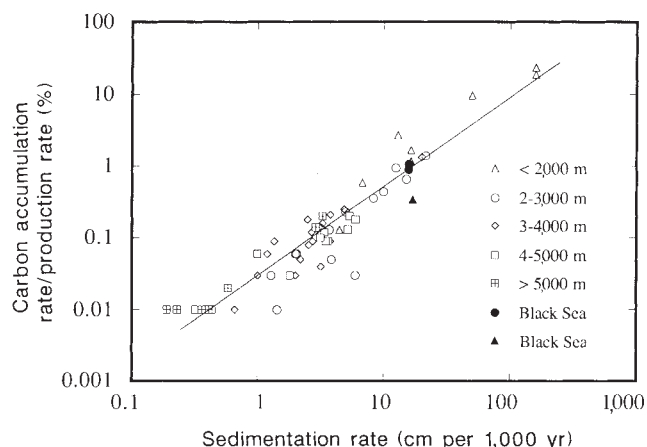


FIG. 3 Organic carbon accumulation rates normalized to primary production rates (% carbon buried) as a function of sedimentation rates. Open symbols represent data from open marine, oxygenated environments and are taken from ref. 29. The solid triangle represents data for core BS4-15, and the solid circles represent data for cores BS4-9 and BS4-14 (Table 1), all normalized to a primary production rate of $200 \text{ g C m}^{-2} \text{ yr}^{-1}$ (ref. 8). Adoption of the primary production value assumed by Deuser⁴ would change the position of the Black Sea data points by a factor of 2, but they would remain within the scatter around the regression line shown. The equation of regression line is $A = 0.028 S^{1.2503}$, where A is the normalized accumulation rate and S the sedimentation rate; the correlation coefficient $r^2 = 0.96$.

continuous deposition of strictly annual lamina couplets in this basin may be incorrect.

The 'burial efficiency' of organic matter in marine sediments (the proportion of the organic carbon accumulating on the sediment surface that becomes permanently buried²⁵) is a strong function of water depth and total sedimentation rate in oxygenated environments⁹. It is widely believed to be higher at equivalent sedimentation rates under anoxic conditions, but this conclusion has been disputed²⁵⁻²⁷. Canfield⁷ used data on the sulphate reduction rates in abyssal sediments from the Black Sea²⁴ to derive the corresponding organic carbon oxidation rates, which yielded very high burial efficiencies of 13-51% and thus supported the conclusion. An alternative way of estimating the degree of preservation of organic carbon in the sediments is to compare the accumulation rates with the primary production rates in the overlying water column. A value for the primary production rate in the Black Sea of $100 \text{ g C m}^{-2} \text{ yr}^{-1}$ (ref. 4) would yield organic carbon burial rates for this basin of 0.7-2.1% (using the data in Table 1). But Sorokin⁸ has argued that the primary production rate in the Black Sea is probably double this value because of the undersampling of the seasonal production maxima¹⁸; measurements of the primary production rate in the Black Sea in May 1988 (the period following the spring bloom) of $0.575 \text{ g C m}^{-2} \text{ d}^{-1}$ (ref. 28) support this argument. Using this new information, the organic carbon burial rates in the deep water sediments of the Black Sea cannot be distinguished from those of oxic open-ocean sediments accumulating at similar water depths and sedimentation rates (Fig. 3). We conclude that organic carbon accumulation rates in anoxic Black Sea sediments have been overestimated in the past because of inaccurate chronologies; contrary to the commonly accepted view, the central and southern Black Sea is not a site of anomalously high organic carbon accumulation, and carbon burial rates are similar to those in the open ocean. □

Received 14 November 1990; accepted 12 March 1991.

1. Woolnough, W. G. *Bull. Am. Ass. Petrol. Geol.* **21**, 1101-1157 (1937).
2. Thiede, J. & van Andel, T. H. *Earth planet. Sci.* **33**, 301-309 (1977).
3. Demaison, G. J. & Moore, G. T. *Org. Geochem.* **2**, 9-31 (1980).
4. Deuser, W. G. *Deep-Sea Res.* **18**, 995-1004 (1971).

5. Calvert, S. E., Vogel, J. S. & Southon, J. R. *Geology* **15**, 918-921 (1987).
6. Hay, B. J. *Paleoceanography* **3**, 491-508 (1988).
7. Canfield, D. E. *Deep-Sea Res.* **36**, 121-138 (1989).
8. Sorokin, Y. I. in *Ecosystems of the World, Vol. 26: Estuaries and Enclosed Seas* (ed. Ketchum, B. H.) (Elsevier, Amsterdam, 1983).
9. Müller, P. J. & Suess, E. *Deep-Sea Res.* **26**, 1347-1362 (1979).
10. Ross, D. A., Degens, E. T. & McIlvaine, J. *Science* **170**, 163-165 (1970).
11. Köglér, F. C. *Meyniana* **13**, 1-7 (1963).
12. Piskaln, C. H. *Trans. Am. Geophys. Un.* **71**, 151 (1990).
13. Slota, P. J., Jull, A. J. T., Linick, T. W. & Toolin, L. J. *Radiocarbon* **29**, 303-306 (1987).
14. Linick, T. W., Jull, A. J. T., Toolin, L. J. & Donahue, D. J. *Radiocarbon* **28**, 522-533 (1986).
15. Vogel, J. S., Southon, J. R., Nelson, D. E. & Brown, T. A. *Nucl. Instrum. Meth.* **233**, 289-293 (1984).
16. Nelson, D. E., Vogel, J. S., Southon, J. R. & Brown, T. A. *Radiocarbon* **28**, 215-222 (1986).
17. Degens, E. T. *et al. Neues Jb. Geol. Paläontol. Mh.* 1980, 65-86 (1980).
18. Honjo, S. *et al. Mitt. Geol. Paläontol. Institut Univ. Hamburg* **62**, 19-39 (1987).
19. Hay, B. J. *et al. Deep-Sea Res.* **37**, 911-928 (1990).
20. Broecker, W. S. & Peng, T.-H. *Tracers in the Sea*, 690 (Eldigio, Palisades, New York, 1982).
21. Murray, J. W., Top, Z. & Ozsoy, E. *Deep-Sea Res.* (in the press).
22. Jones, G. A. *Trans. Am. Geophys. Un.* **71**, 152 (1990).
23. Crusius, J. & Anderson, R. *Paleoceanography* (submitted).
24. Vaynshteyn, M. B., Tokarev, V. G., Shakola, V. A., Lein, A. Y. & Ivanov, M. V. *Geokhimiya* **7**, 1032-1044 (1985); (Engl. transl.) 110-122.
25. Henrichs, S. M. & Reeburgh, W. S. *Geomicrobiol. J.* **5**, 191-237 (1987).
26. Pedersen, T. F. & Calvert, S. E. *Bull. Am. Ass. Petrol. Geol.* **74**, 454-466 (1990).
27. Calvert, S. E. & Pedersen, T. F. in *Productivity, Accumulation and Preservation of Organic Matter in Recent and Ancient Sediments* (eds Whelan, J. K. & Farrington, J. W.) (Columbia University Press, in the press).
28. Karl, D. M. & Knauer, G. A. *Deep-Sea Res.* (in the press).
29. Sarnthein, M., Winn, K., Duplessy, J.-C. & Fontugne, M. R. *Paleoceanography* **3**, 361-399 (1988).

ACKNOWLEDGEMENTS. Participation in the cruise of RV Knorr was supported by the Natural Sciences and Engineering Research Council of Canada (S.E.C.) and the NSF (R.E.K.). The help of cruise participants in retrieving the cores is gratefully acknowledged. Support for the McMaster University and University of Arizona AMS facilities was provided by the NSERC and NSF, respectively.

Prospects for eruption prediction in near real-time

Barry Voight* & Reinold R. Cornelius

College of Earth and Mineral Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA

* Also at US Geological Survey, Igneous & Geothermal Processes, 345 Middlefield Road, Menlo Park, California 94025, USA

THE 'materials science' method for eruption prediction¹⁻³ arises from the application of a general law governing the failure of materials: $\Omega^{-\alpha} \dot{\Omega} - A = 0$, where A and α are empirical constants, and Ω is an observable quantity such as ground deformation, seismicity or gas emission. This law leads to the idea of the 'inverse-rate' plot, in which the time of failure can be estimated by extrapolation of the curve of $\dot{\Omega}^{-1}$ versus time to a pre-determined intercept. Here we suggest that this method can be combined with real-time seismic amplitude monitoring to provide a tool for near-real-time eruption prediction, and we demonstrate how it might have been used to predict two dome-growth episodes at Mount St Helens volcano in 1985 and 1986, and two explosive eruptions at Redoubt volcano in 1989-90.

The materials science method may be applied to data sets from appropriately designed and located geodetic monitoring systems, for example from repeated electronic distance or tilt measurements. The method should be most useful, however, for

seismic data, because seismicity is the phenomenon most commonly monitored to predict volcanic eruptions^{1,4-10}. Characteristics of the seismic activity of volcanoes include the number of earthquakes, their energy and moment, and the seismic strain release. The number of earthquakes is less fundamental than their energy content.

Seismic energy calculations conventionally require study of records, recognition and classification of discrete earthquakes, and application of correlations involving energy-magnitude or energy-coda relations. Some of these operations may be computerized, but the conventional approach is manual and time-consuming, and energy calculations may lag the actual events by hours or days. Although useful, prediction on the basis of seismic energy is frequently not real-time.

During periods of intense seismicity, earthquake traces may overlap or be superposed on tremor of significant amplitude; individual earthquakes cannot be accurately counted, and coda lengths cannot be measured. Use of more distant seismic stations minimizes the noise due to tremor and small earthquakes, but introduces bias towards large-magnitude events and systematic error into estimates of coda magnitude¹¹.

These problems can be minimized by continuous measurement of the mean amplitude of seismicity¹¹⁻¹⁵. The real-time seismic amplitude measurement (RSAM) system used by the US Geological Survey provides information from up to eight telemetered seismic stations¹⁶. The absolute voltage (amplitude) level for each station is digitized, averaged and appended in near real-time (averaging is over 1- or 10-min intervals) to a

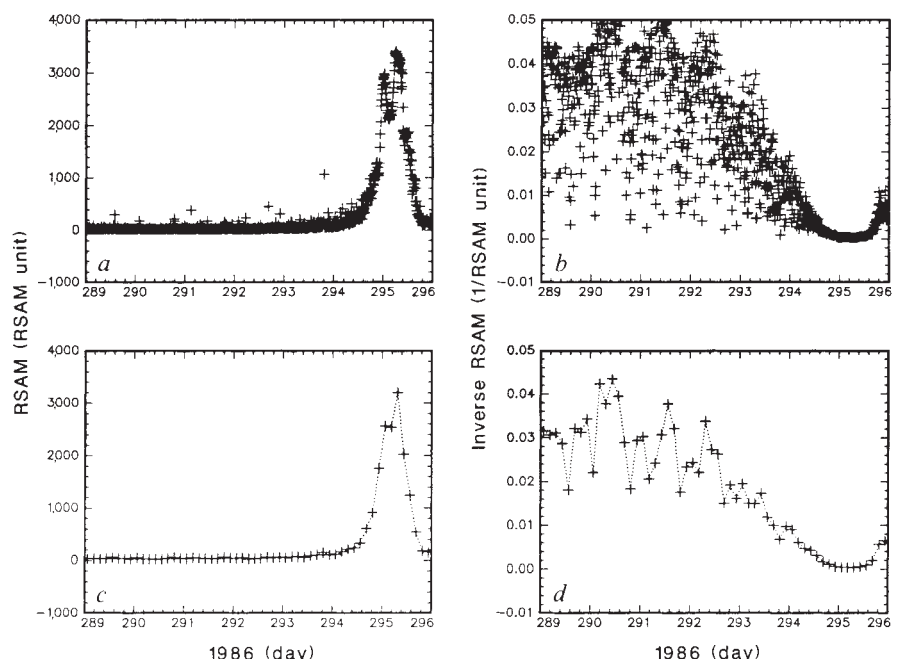
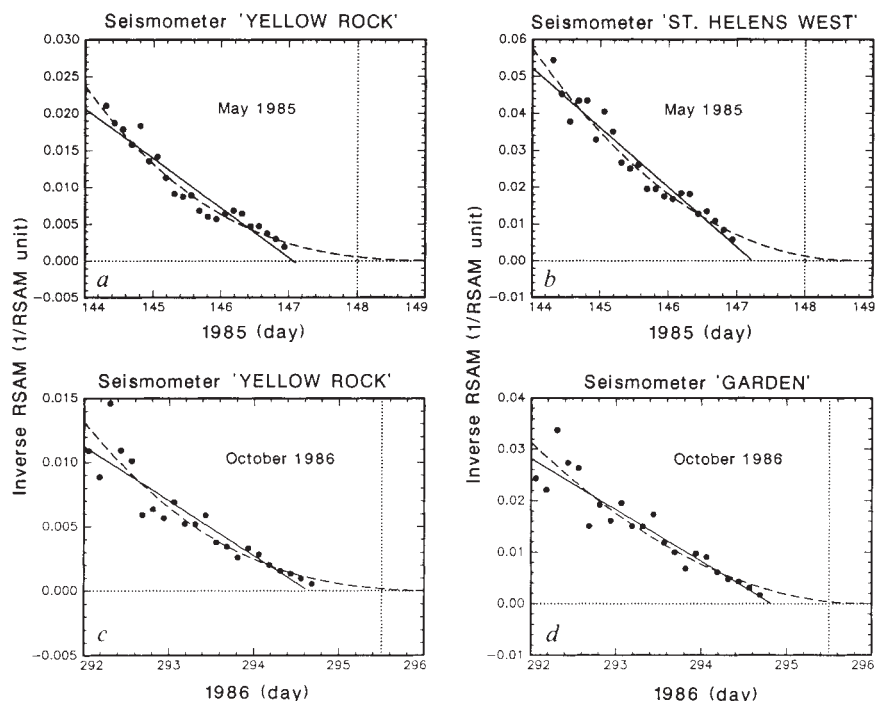


FIG. 1 RSAM from seismometer 'Garden', Mount St Helens, between 16 October (day 289) and 22 October 1986 (day 296). Time used is GMT. *a*, RSAM averaged over periods of 10 min. *b*, Inverse RSAM calculated for *a*. *c*, RSAM averaged over periods of 3 h. *d*, Inverse-RSAM calculated for *c*.

FIG. 2 Inverse-RSAM from three stations for the May 1985 and October 1986 eruptions at Mount St Helens. Data are averaged over 3 h. Station Yellow Rock is in the crater, 1 km from the dome; Garden, 500 m north of base of dome; St Helens West, 4 km outside the crater. Time used is GMT. *a, b*, Data from 1985 are for Yellow Rock and St Helens West between 24 May (day 144.3) and 27 May (day 147.0). *c, d*, Data from 1986 for Yellow Rock and Garden between 19 October (day 292.0) and 21 October (day 294.7). Data sets for the same eruption show similar features. Solid lines are fits to the data with $\alpha=2$; dashed lines are fits by numerical methods with $1 < \alpha < 2$. Data sets are terminated roughly at the time of the point of inflection of the RSAM curve. Vertical lines represent approximate time of observed onset of surface events (pervasive dome disruption) and/or eruption.



data file on a multi-user computer system. A recent modification of the seismic-acquisition system used to monitor Redoubt Volcano, Alaska, allows average amplitudes in each of 16 spectral bands to be computed from digitized analog signals with a fast fourier transform (FFT)¹⁶. This has been termed SSAM (seismic spectral amplitude measurement)¹⁷. Such measurements, conventional or spectrally bracketed, provide a refined means of defining the relative intensities of eruptive episodes regardless of the type of activity (discrete earthquake, closely-spaced or overlapping earthquakes, or tremor).

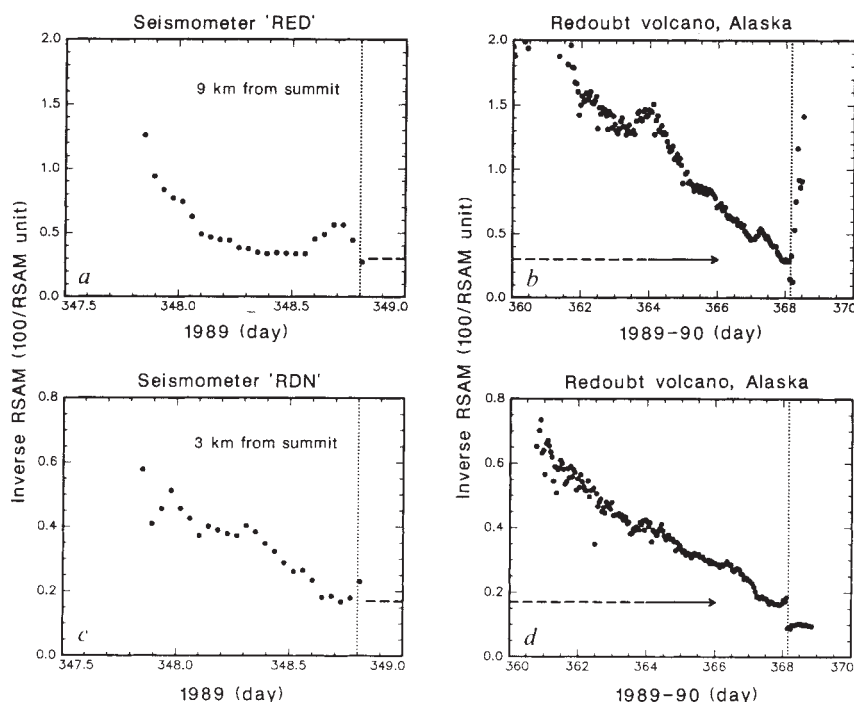
If transducer output is known accurately, then each 1-min average can be scaled to give average ground displacement. This measurement can be adjusted for geometric spreading and instrument amplification to obtain the 'reduced displacement'¹⁸.

Thus RSAM is directly proportional to commonly measured earthquake parameters such as moment rate or energy rate, as these also are proportional to reduced displacement.

Even if the proportionality constants are not precisely known, in terms of the materials science prediction method the RSAM output is directly analogous to the rate of energy or moment. The procedure of plotting inverse rate ($\dot{\Omega}^{-1}$) against time (t) may therefore be applied simply by inverting RSAM without further data manipulation.

Here we present examples for two distinctive styles of eruption involving Mount St Helens, Washington, in 1985 and 1986, and Redoubt volcano, Alaska, in 1989–90. Seismic activity associated with three growth episodes of the composite dacite dome at Mount St Helens produced very similar precursory RSAM

FIG. 3 Inverse-RSAM for stations RED (upper) and RDN (lower) at Redoubt volcano, before two major eruptions in 1989–90. Time used is GMT. Station RDN used automatic gain control, which may have influenced the measurements. *a, c*, The 1989 explosion of 14 December at 10:13 Alaskan Standard Time, AST (day 348.80), which resulted in an eruption plume 33,000 feet high. *b, d*, The 2 January 1990 vent-clearing eruption, which consisted of two explosive events at 17:48 and 19:27 AST (day 368.12 and 368.19); each resulted in a plume 40,000 feet high. Times of eruptions are marked as vertical dotted lines. A finite 'failure rate' (horizontal arrows) could have been estimated from plots for the 14 December eruption and the intercept with the inverse-RSAM curve in January used to predict time of eruption.



curves^{11,19}, with curvature typically $1 < \alpha < 2$, consistent with the materials science theory. The RSAM change is not exponential as previously assumed¹¹ (exponential growth implies $\alpha = 1$, but the best fits to the RSAM curves have $\alpha \neq 1$).

The last of these dome-building events involved endogenous growth and lava extrusion on 22 October 1986 (day 295)¹⁹. Figure 1 shows RSAM growth curves for both 10-min and 3-h average amplitudes monitored at one station near the dome. Also shown are the inverse-RSAM curves, a cornerstone of the materials science approach. Failure is indicated when a critical inverse value Ω_r^{-1} is obtained; if Ω_r^{-1} is uncertain, an approximate upper-bound failure date can be obtained from the point of intersection of the extrapolated inverse-rate curve with the time axis (this singularity occurs for $\alpha > 1$).

Figure 1 illustrates the reduction in scatter when amplitudes are averaged for periods exceeding an hour. Graphical extrapolations of inverse-RSAM before day 292 indicated that an event should be expected after day 296, but the processes associated with RSAM accelerated after day 292 leading to an earlier prediction date. Subsets of inverse-RSAM data after day 292 and similar data for the May 1985 eruption (Fig. 2) demonstrate the consistency of the method as observed over a broad network of seismic stations. Linear (assumed $\alpha = 2$) and curved ($1.2 < \alpha < 1.6$) extrapolations are shown; although these curves are perturbed by the small RSAM shift one day before eruption, the predictive capability of the inverse-RSAM procedure seems obvious. The onset of precursory RSAM acceleration varied at Mount St Helens from ~5 days before the October 1986 event (Fig. 1) to ~12 and 7 days, respectively, before the May 1985 and May 1986 events.

In practice, graphical inspection of the inverse-RSAM pattern as it approaches the abscissa is more useful than trying to predict events from the empirical constants. For $\alpha < 2$, the inverse-rate curves have a concave upward bend; linear extrapolation thus yields a lower-bound prediction.

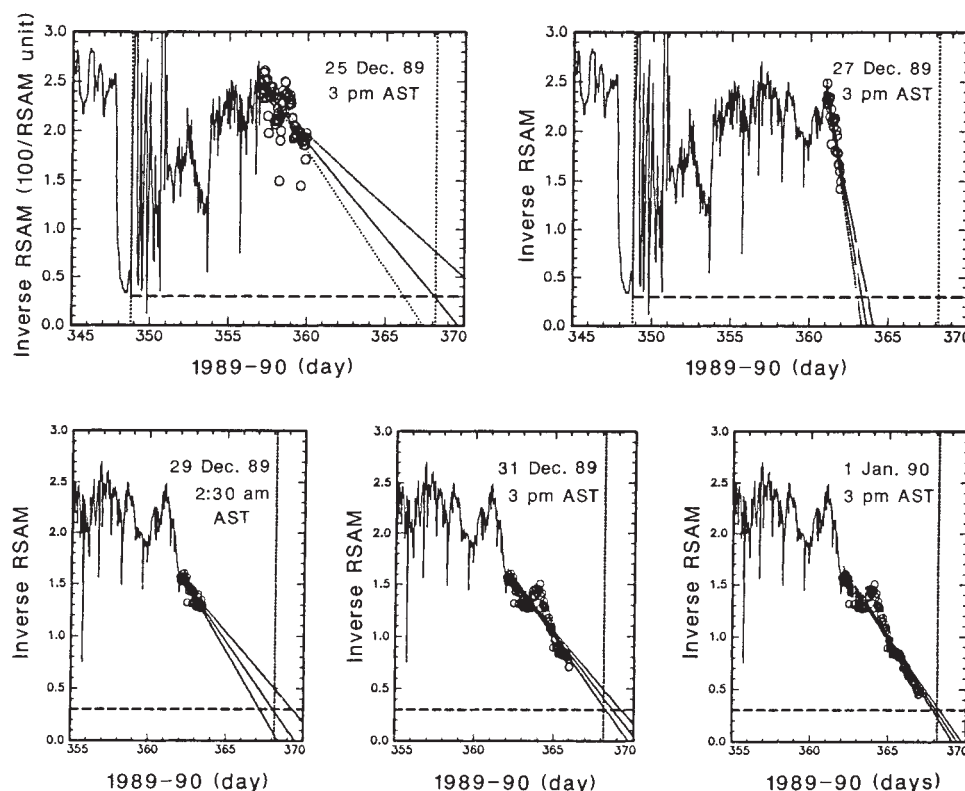
On 14 December 1989, Redoubt volcano (Alaska) erupted after 23 years of quiescence²⁰. Seismic activity was tracked by the RSAM system to provide consecutive 10-min averages of the absolute amplitude. Spectral analysis¹⁷ showed that the main

seismic excitation was in the range 1–2.5 Hz. For the largest eruptions the characteristic signals were observed 20 km from the summit; for the smaller events, near-source seismographs were required. Clear RSAM signals were seen before six of the eight significant eruptions that occurred between 5 March and 21 April, when at least one seismograph within 3 km of the summit was operating continuously.

Accelerations of RSAM amplitude also preceded the powerful vent-clearing phreatomagmatic explosions of 14–22 December and the dome-destroying explosions of 2 January AST (day 368). The 2 January explosions triggered a series of pyroclastic flows which in turn generated hot lahars and hyperconcentrated floods²⁰. Precursory changes in RSAM for the 14 December event developed over little more than one day (Fig. 3a and c), whereas the pattern before 30 December (day 364) warns of eruption several days in advance. The inverse-RSAM 'failure rate' (Ω_r^{-1}) experienced in December (Fig. 3a and c) provides a reasonable guide to predict eruption on 2 January (note intersection of horizontal arrow with inverse-RSAM curves, Fig. 3b and d). This procedure is an alternative to the extrapolation of inverse-RSAM to the time abscissa (which would imply infinite rate), although it must be recognized that precursory processes (and hence critical RSAM values) may vary from event to event. Further study of this important issue is required.

Obviously, it is easier to recognize trends in hindsight with data sets complete to the date of eruption (Fig. 3), than it is to interpret real-time data as it arrives. As an example, we reconstruct in Fig. 4 sequential 'snapshots' of inverse-RSAM data, averaged hourly, before the 2 January eruption. We show analyses for five successive times between 25 December and 1 January, using the assumption $\alpha = 2$ (linear extrapolation) with the latest data (circles). These analyses have been produced in hindsight, as the method was not used at Redoubt in 1989. In Fig. 4, inverse-RSAM extrapolations after 29 December are all consistent with eruption on 2 January. Analysis for 27 December warned of the possibility of an early eruption, but from a trend shift that occurred shortly afterwards, we could have revised the interpretation.

FIG. 4 Sequential analyses from station RED before eruptions on 2 January 1990. This sequence simulates analyses which might have been made as the time of eruption approached, although no inverse-rate analyses were actually performed in 1989. The eruptions on 14 December and 2 January are shown as vertical dotted lines; finite 'failure rate', as estimated from the December eruption, is shown as horizontal dashed line. Data are averaged over 1 h. Data used for analyses are shown as circles, other data as continuous lines. Analyses consist of linear least squares fits, assuming $\alpha = 2$, with an error envelope at the 97.5% confidence level. Bounds are set on the time of eruption by the intersection of linear projections and their envelopes with the estimated finite 'failure rate' (lower bound) and/or a possible infinite rate (upper bound). Analysis on 27 December (six days before eruption) might have led to a false early warning, but the trend changed shortly afterwards, and subsequent analyses suggested fairly accurate eruption times.



The examples presented here indicate that the method could provide a useful guide to prediction of eruptions, although the reliability of the method may be influenced by the regularity of structural mechanisms and loading conditions. Sharp breaks in observed rate with time, such as for Redoubt volcano (Fig. 4), may warn of changes in loading or structure. If the data imply RSAM change that is not (at least approximately) smooth and continuous, segmented analysis may be necessary. In such an approach, the initial value should be reassigned to exclude portions of the data set no longer considered applicable (Figs. 2 and 4). To apply the theory of this paper blindly, irrespective of such indications of change, is to misapply it.

A limitation of the method is the incorporation of seismic

noise into the amplitude calculations, but misinterpretation can be minimized by comparing RSAM with (for example) gas, deformation or meteorological information, plotted on the same time base, and by studying analog seismic records or spectral data to identify tectonic earthquakes, fumarolic noise, storm noise and kindred phenomena¹¹. We recognize that this method, in common with all known eruption prediction methods, will not necessarily work for all volcanoes, under all circumstances^{3,10}; of 23 small explosive events at Tokachi-dake, Japan in 1988–89, only a third had common immediate precursors¹⁵. Predictive capability is best achieved by using a combination of data sets and methods, rather than by reliance on any single procedure. □

Received 9 October 1990; accepted 8 March 1991.

1. Voight, B. *Nature* **332**, 125–130 (1988).
2. Voight, B. *Science* **243**, 200–203 (1989).
3. Tilling, R. I. *Nature* **332**, 108–109 (1988).
4. Tokarev, P. I. *Bull. Volcan.* **35**, 243–250 (1971).
5. Tokarev, P. I. in *Forecasting Volcanic Events* (eds Tazieff, H. & Sabroux, J.-C.) Ch. 9 (Elsevier, Amsterdam, 1983).
6. Shimozuru, D. in *The Surveillance and Prediction of Volcanic Activity*, 19–45 (UNESCO, Paris, 1972).
7. Minakami, T. in *Physical Volcanology* (eds Civetta, L., Gasparini, P., Luongo, G. & Rapallo, A.) 313–333 (Elsevier, Amsterdam, 1974).
8. Malone, S. et al. *Science* **221**, 1376–1378 (1983).
9. Swanson, D. A. et al. *J. Geodynam.* **3**, 397–423 (1985).
10. Decker, R. W. *Ann. Rev. Earth planet. Sci.* **14**, 267–291 (1986).
11. Endo, E. T. & Murray, T. *Bull. Volcanology* (in the press).
12. Sassa, K. *Mem. Coll. Sci. Kyoto Univ.* **19**, 11–56 (1936).

13. Schick, R. et al. *J. Volcan. geotherm. Res.* **14**, 261–279 (1982).
14. Nishi, K. *Annls Disaster Prev. Res. Inst. Japan* **B 30**, 1–18 (1990).
15. Okada, H. *Bull. Volcan. Soc. Japan* **35**, 175–203 (1990).
16. Murray, T. & Endo, E. T. *US geol. Surv. Open File Rep.* 89–684 (1989).
17. Stephens, C. D. et al. *EOS* (in the press).
18. Fehler, M. J. *geophys. Res.* **88**, 3476–3484 (1983).
19. Endo, E. T., Dzulisin, D. & Swanson, D. A. in *Magma Transport and Storage* (ed. Ryan, M. P.) Ch. 15 (Wiley, Chichester, 1990).
20. Alaska Volcano Observatory *EOS* **71**, 265–275 (1990).

ACKNOWLEDGEMENTS. We thank colleagues and US Geological Survey staff for assistance, and S. Malone, T. Murray, J. Power and R. I. Tilling for constructive criticism of the manuscript. This work was conducted under a cooperative project involving Penn State University, the Cascade Volcano Observatory (US Geological Survey) and the Alaska Volcano Observatory (US Geological Survey, University of Alaska Geophysical Institute and the Alaskan Division of Geological and Geophysical Surveys). Support was provided by NSF.

A fossil smut fungus from the anthers of an Eocene angiosperm

R. S. Currah & R. A. Stockey

Department of Botany, University of Alberta, Edmonton, Alberta, Canada T6G 2E9

WHILE studying the floral morphology of a permineralized angiosperm from a Middle Eocene chert near Princeton, British Columbia, we encountered well-preserved masses of fungal spores in otherwise apparently normal anthers. The spore masses, which replaced the pollen, consist of single-celled, irregularly shaped to ovoid, minutely pitted spores. On the basis of intact nature of infected anthers and size, ornamentation and disposition of spores within anther locules, we have identified this material as an anthericolous smut fungus. This unequivocal fossil of the Ustilaginales shows that at least 48 million years ago the order had representatives with anthericolous sori, a characteristic of some extant smut taxa that permits effective insect dispersal of the fungus among populations of specific host plants.

The Ustilaginales includes about 1,000 species of obligate parasites that attack over 75 families of flowering plants, particularly the graminoid families, Cyperaceae and Poaceae¹. As biotrophic parasites, they draw nutrients from plants without seriously affecting the reproductive potential of host species. Smut fungi reproduce in their angiosperm hosts by forming masses of thick-walled teliospores in sori which develop within the tissues and beneath the epidermis of leaves, stems, roots, ovaries, anthers and so on. Rupture of overlying epidermal tissue with exposure of a dusty excrescence of darkly pigmented spores is often the only indication that a plant is infected.

About 80 small flowers 0.9-mm long and 0.8 mm in diameter have been found in the Princeton chert (Fig. 1a and c). Twelve were examined for pollen by scanning electron microscopy. All specimens show masses of fungal spores, about $8 \times 6 \mu\text{m}$ (Fig. 1b) and lack pollen in the anthers. The bilateral symmetry and the broken bases of the flowers suggest that they are borne on

spikes. They possess a small receptacle, no disc, one to two perianth segments, four basifixed anthers opening by longitudinal slits, stamens free from corolla, non-connate filaments, four free carpels, a unilocular ovary, one ovule per locule and a single-seeded fruit with an aerenchymatous endocarp^{2,3}. These characters suggest affinities with one of several aquatic monocotyledonous families in the Alismatidae, particularly the Aponogetonaceae, a small family native to the Old World Tropics, Australia and South Africa^{4,5}. The fossil flowers, however, have four stamens rather than the six or more of the Aponogetonaceae. Pollen morphology, one of the most important characters used in identification of fossil flowers, is lacking because of the presence of teliospores. Formal taxonomic description of these flowers will be published elsewhere.

Extant smut taxa of Alismatidae are included in the genera *Doassansia*, *Burrillia* and *Tracya*⁶. These differ from the fossil material in producing teliospores in spore balls that develop in sori within leaves, petioles and stems. There are a few records of anthericolous smuts in extant Liliaceae⁷, but none apparently for extant alismatidaceous taxa.

The presence of systemic biotrophic fungi in some angiosperms can cause abnormal developmental processes in flowers⁸. The effects of the smut fungus on floral development may explain the incongruous number of stamens in our fossil material. For example, dioecious plants of *Silene* infected with *Microbotryum violaceum* often seem to have perfect flowers because the smut causes the staminodes of female flowers to develop into stamens, the anthers of which contain teliospores rather than pollen. Smutty anthers are reported to be as attractive to pollinators as normal anthers, so anthericolous smuts are assured of dispersal among their particular host species⁹. Identification of extant species and genera within the Ustilaginales is based on colour, size and shape of spores and spore balls, spore wall ornamentation, which varies from smooth to pitted or reticulate (Fig. 1d) to verrucose, the identity of the host plant and the mode of formation of basidiospores, which develop directly on a germ tube arising from a germinating teliospore. Without observations on the mode of basidiospore formation, it may be impossible to place the fossil smut in a modern genus. But until recently,

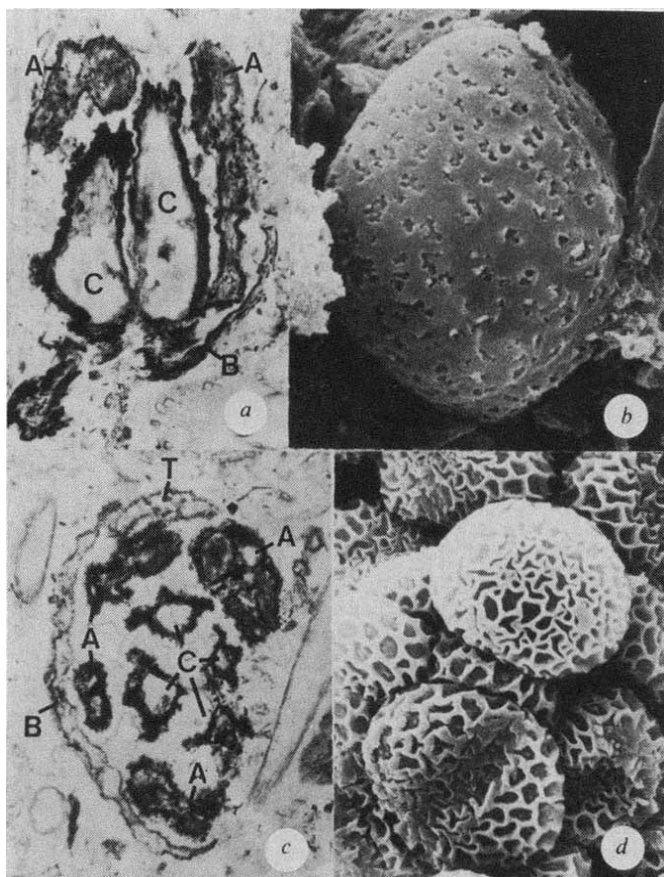


FIG. 1 a, A longitudinal section showing carpels and stamens of the permineralized Eocene flower. A, anther in which the smut sorus has replaced the pollen. B, bract. C, carpel. Magnification, $\times 42$. b, Scanning electron microscopy of an individual, single-celled, minutely pitted, fossil teliospore within the sorus. Magnification, $\times 5,780$. c, Transverse section of the flower. A, anther. B, bract. C, carpel. T, tepal. Magnification, $\times 60$. d, Scanning electron microscopy of pitted teliospores within a sorus on an ovary of *Scorzonera humilis* L. infected with the extant smut *Ustilago scorzonerae* (Albertini and Schweinitz) Schroeter. Magnification, $\times 2,720$.

most smuts forming simple, small (4–18- μm diameter) single-spored teliospores in the reproductive organs of angiosperms were placed in the genus *Ustilago*^{1,10}. The genus has now been divided into a number of smaller genera⁶. Of these, the genus *Microbotryum* is the closest to the fossil representative, with respect to teliospore morphology and sorus location. *Microbotryum violaceum* (Persoon: Persoon) Deml and Oberwinkler (= *Ustilago violacea* (Pers.: Pers.) Roussel) forms sori of single, globose to somewhat irregular, pale lilac, finely and irregularly reticulate spores within the anthers of Caryophyllaceae.

The origin and evolution of the Ustilaginales are problematic because of the absence of any confirmed fossil material¹¹. The few references to their fossil record are based on questionable interpretations of what might be a mixed collection of isolated teliospores from Ustilaginales and Uredinales in Tertiary lignites^{11,12}. On the basis of the intact antheridial sori and the morphology of the teliospores, our material represents fossils of the Ustilaginales and indicates that the highly specialized dispersal mechanism of some modern representatives of the order had evolved by at least 48 million years ago. □

Received 11 December 1990; accepted 12 February 1991.

1. Fischer, G. W. *Manual of North American Smut Fungi* (Ronald, New York, 1953).
2. Cevallos-Ferriz, S. R. S. thesis, University of Alberta (1987).
3. Cevallos-Ferriz, S. R. S. & Stockey, R. A. *Rev. Palaeobot. Palynol.* (in the press).

4. Cronquist, A. *An Integrated System of Classification of Flowering Plants* (Columbia University Press, 1981).
5. Lawrence, G. H. M. *Taxonomy of Vascular Plants* (Macmillan, New York, 1952).
6. Vanky, K. *Cryptogamic Studies* **1**, 1–159 (1987).
7. Lindeberg, B. *Symb. Bot. Upsal.* **16**, 1–175 (1959).
8. Clay, K. in *Coevolution of Fungi with Plants and Animals* (eds Pirozynski, K. A. & Hawksworth, D. L.) 79–105 (Academic, 1988).
9. Jennerston, O. *Oikos* **51**, 163–170 (1988).
10. Duran, R. in *The Fungi* (eds Ainsworth, G. C., et al) 281–300 (Academic, 1973).
11. Pirozynski, K. A. *Rev. Phytopath.* **14**, 237–246 (1976).
12. Tiffney, B. H. *Occasional Pap. Farlow Herbarium Cryptogam. Bot.* **7**, 1–42 (1974).

Perceptual filling in of artificially induced scotomas in human vision

V. S. Ramachandran & R. L. Gregory*

Psychology Department 0109, University of California, San Diego, La Jolla, California 92092-0109, USA

* Department of Psychology, University of Bristol, Bristol, UK

PATIENTS with scotomas or blind-spots in their visual field^{1–5} resulting from damage to the visual pathways often report that the pattern from the rest of the visual field ‘fills in’ to occupy the scotoma. Here we describe a novel technique for generating an artificial perceptual scotoma which enabled us to study the spatial and temporal characteristics of this filling-in process. A homogeneous grey square subtending 1.5° was displayed against a background of twinkling two-dimensional noise of equal mean luminance (Fig. 1). On steady eccentric fixation for 10 s the square vanished and was filled in by the twinkling noise from the surround. Using this display we found that ‘filling in’ is an active visual process that probably involves creating an actual neural representation of the surround rather than merely ignoring the absence of information from the scotoma; filling in can occur separately for colour and texture, suggesting separate mechanisms; the filling-in process does not completely suppress information from the scotoma, even after an image has faded completely from consciousness it can nevertheless contribute to motion perception; and the process can be strongly influenced by illusory contours.

Patients with discrete lesions in their retinas or in primary visual cortex usually have a ‘blind spot’ or scotoma in their visual field as measured by conventional perimetry^{1–5}. Oddly enough, if the patient gazes at a companion seen against a background of wallpaper, the companion’s head may disappear but the gap gets ‘filled in’ by the surrounding wallpaper. It is said that King Charles II used to decapitate his courtiers using this benign procedure, although he used his natural blind spot rather than a scotoma.

Of the natural blind spot (corresponding to the optic nerve head), Sir David Brewster⁶ has written: “We should expect, whether we use one or both eyes, to see a black or dark spot upon every landscape within 15° of the point which most particularly attracts our notice. The Divine Artificer, however, has not left his work thus imperfect. . . the spot, in place of being black, has always the same colour as the ground.”

We know surprisingly little about the nature of the neural representation that corresponds to the filling in of scotomas and blind spots. To study the spatial and temporal characteristics of this process we created a novel stimulus that results in a temporary reversible scotoma in normal individuals^{7,8}. A homogeneous grey square subtending $1.5^\circ \times 1.5^\circ$ was displayed against a background of dynamic twinkling noise and the subjects were asked to fixate steadily on a spot 6° to the right (or left) of the square (Fig. 1). The twinkle had the same mean luminance (25 cd m^{-2}) as the square. Four naive subjects were used in all the experiments described here. After a brief period of steady fixation they reported that the square vanished completely and was filled in by the surrounding twinkle. The average

time for fading and filling in was 5.04 s ($n = 40$; four subjects, 10 trials each.)

Does this filling-in process involve creating an actual neural representation of the 'twinkle' in the brain areas to which the grey square projects, or does it merely involve ignoring the absence of information? (After all, one is unaware of the huge gaping hole behind one's head, but one would not want to conclude, for example, that the wallpaper had been filled in behind the head.) One hint comes from the fact that the filling-in process often appeared gradual; after the square's borders faded, the twinkle filled in from outside to inside slowly, taking 2 or 3 s. This suggests an active process.

Is filling-in a unitary process or can it occur separately for colour and texture? To find out, we modified Fig. 1 so that the surround consisted of sparse black dots twinkling on a homogeneous pink background, and the square consisted of black spots moving horizontally against a grey background (Fig. 2) that was equiluminous with the pink. This square, defined by a difference in motion as well as colour, also faded completely in a few seconds, but we found that the filling-in occurred in two distinct stages. First, the homogeneous grey region in the square vanished and was filled in by the pink from the surround, so that one now had the experience of seeing the black dots moving against the filled-in pink that did not actually exist on the retina. Once this had occurred, the moving spots also faded and were replaced a few seconds later by the twinkling spots in the surround. Thus, there may be separate fill mechanisms for colour and texture corresponding, perhaps, to the different extrastriate visual areas⁹⁻¹² that are thought to be specialized for colour or motion (twinkle). Perhaps the colour border fades first and the surround colour is then assigned to this region (in the 'colour areas' such as V4). Because the moving spots have not yet faded, the visual system 'assumes' that the spots must be moving against the filled-in pink.

Does the fading of the square (and subsequent filling in) depend mainly on the adaptation of neural detectors that extract borders of the square or of the homogeneous region inside it?

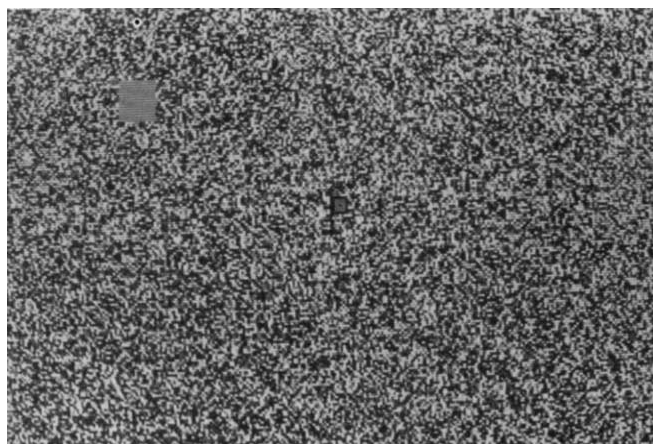


FIG. 1 Stimulus used to produce an artificial perceptual scotoma. The background consisted of twinkling spots of eight different grey levels. The square subtended $1.5^\circ \times 1.5^\circ$ and it had the same mean luminance (50 cd m^{-2}) as the twinkling texture. The fixation spot was about 6° away from the border of the square. On steady fixation the square vanished in about 5 s and was filled in by the twinkling noise in the surround. A similar fading and filling in of texture was originally described by Ramachandran and Anstis⁷ and Ramachandran⁸ but in their stimulus the square was a 'window' filled with horizontally moving dots rather than a homogeneous grey. Also, we found that if the square was very small ($<0.2^\circ$) it could be seen to vanish even if it was very bright or dark, that is non-equiluminous with the surround. (The effect could then be seen even if the fixation spot was only 2° from the square.) The fading occurred even more quickly ($<2 \text{ s}$) if the square was in a different stereoscopic plane (nearer or further) than the twinkling texture. This was confirmed by all four subjects.

To answer this question we allowed the filling in to occur and then suddenly switched off and replaced the square with a slightly smaller concentric square ($1.3^\circ \times 1.3^\circ$). Because the region corresponding to this small square has already been filled in with twinkle, does one have to start the filling in again? The answer is yes. The small square did become clearly visible and took several seconds to fade and become 'filled in' again (mean = 2.29 s). Thus the fading of contours is critical to the whole process.

How deep is the suppression of information within the scotoma? To investigate this we waited until the square had faded completely, then switched off the square and replaced it with an identical square that was shifted by about 0.4° . To our surprise, we found that the second square seems to 'move' or 'arrive' at its location, instead of just appearing out of nowhere. Even though the first square was not consciously visible it could nevertheless provide an input to apparent motion perception. A more formal experiment was conducted on the four subjects using a smaller square ($1^\circ \times 1^\circ$). On each trial we waited until the subject reported complete fading, then we switched off the square and replaced it with a new square that was either to the left or to the right of the original. Left-right trials were presented in random order and the subject's task was to simply report the direction. Performance on this task was well above chance (75 out of 80 trials; four subjects, 20 trials each). Furthermore, it was clear from the subjects' phenomenological reports that they were actually experiencing a sensation of motion and not simply reporting a change of location. We postulate, therefore, that the fading of the square from consciousness occurs from area V4, the dorsolateral area (DL) and other 'form' areas, but that information about its motion is processed separately in the middle temporal area (MT).

Our next experiment also explored the depth of suppression within the scotoma. We waited until the square was replaced with twinkle, then we suddenly introduced a small (0.2°) red test-probe in the centre of the grey square. All four subjects reported that the spot instantly 'overcame' the filling in, but only in the small region corresponding to the spot, so that they saw a red spot surrounded by the filled-in twinkle that did not

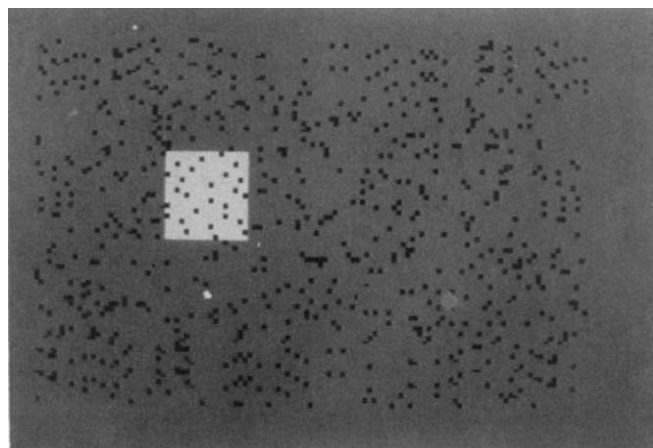


FIG. 2 Schematic illustration of stimulus used to demonstrate independent filling-in mechanisms for colour and twinkling texture. The homogeneous grey square subtended $2.6^\circ \times 2.8^\circ$. The black spots in the square subtended 5 minutes of arc and moved horizontally as in a 'conveyor belt'. The surround consisted of randomly flickering ('twinkling') black spots on a homogeneous pink background (depicted dark grey) whose mean luminance was the same (80 cd m^{-2}) as the grey square. The background subtended $19^\circ \times 15^\circ$. Four subjects were asked to view this display and asked whether the twinkle and colour both filled in at the same time, the twinkle filled in first, or the colour filled in first. On most trials the subjects reported that the colour filled in first (mean = 2.83 s; $n = 40$) followed a few seconds later by the twinkling noise (mean = 4.05 s).

actually exist around it on the retina. That the red spot is seen to be actually surrounded by the filled-in twinkle is additional evidence for an active filling-in process. If no such process exists, one might have simply experienced a vague inability to decide what was immediately around the red spot.

Next, we explored the temporal characteristics of the filling-in process using Fig. 1. After filling in had occurred, we switched off the entire display and replaced it with a homogeneous grey field of equal luminance to the square (the square subtended $1.3^\circ \times 1.3^\circ$). We now found that there was a square patch of twinkly noise in the area that was originally occupied by the square, and this patch persisted for as long as 2–3 s (mean for four subjects = 2.44 s; $n = 40$). We conclude that a time-varying neural representation of the surrounding twinkle is created within the scotoma and perhaps this representation persists for 2–3 s after the display is switched off. An alternative possibility is that the twinkle creates some peculiar state of adaptation in the surround, and this adaptation subsequently 'induces' a spatially heterogeneous twinkle in the region corresponding to the square. Even so, whatever mechanism is responsible for this induction of twinkle in the region originally occupied by the square is unlikely to be very different from the process causing the filling in of the scotoma in the first place. What is truly remarkable about the percept is its dynamic nature, that it is actually seen to twinkle within the scotoma. When we changed the temporal frequency of the surrounding twinkle to about 10 Hz, from 30 Hz, we found that the perceived temporal frequency of the persistent twinkling patch was also reduced correspondingly.

Is the filling-in process specific to the pattern in the surround? We tried to answer this by using displays such as Fig. 3a, in which the surround was a diagonal twinkling ('counterphase

flickering') square wave grating. We waited for fill to occur, then replaced the background lines with twinkling two-dimensional noise. Subjects typically reported that this procedure restored the grey square momentarily (1 s) and that it had to be filled in again by the new twinkling texture in the surround. Scotomas can also be produced, although less vividly, using stationary patterns such as Fig. 3b and c. In Fig. 3b the surround consists of a stationary horizontal square wave grating, and in Fig. 3c it was composed of ordinary type-written English, Latin, or 'non-sense' text. The filling in of text was especially striking and was reported by all four subjects, although, needless to say, none of them could actually read the text within the filled-in region.

Equiluminous chromatic borders^{13,14} and luminance edges¹⁵ also tend to fade during optical image stabilization or steady fixation¹². We found that this was true even if an achromatic (grey) square was displayed on an equiluminous yellow background. The square (subtending 2°) faded in about 3–4 s and was replaced by the yellow from the background. The presence of the yellow within the filled-in region cannot be explained in terms of conventional colour contrast or adaptation effects. Again, if a black spot (0.3° diameter) was introduced in the centre of the square, the filling in of yellow was observed right up to, but not beyond, the edges of the black spot. Yet, if we had a thin black ring (0.3° outer diameter) instead of a spot, the yellow filled the interior of the ring as well, its spread was not 'blocked' by the ring. So the word 'paint' might be a more appropriate metaphor than 'filling in'.

Our next experiment demonstrates an intriguing interaction between illusory contours^{16–22} and the filling-in mechanism. The display consisted of a circular grey disk that straddled the vertical border between two equiluminous regions of different colour (pink and yellow). A vertical illusory contour was then

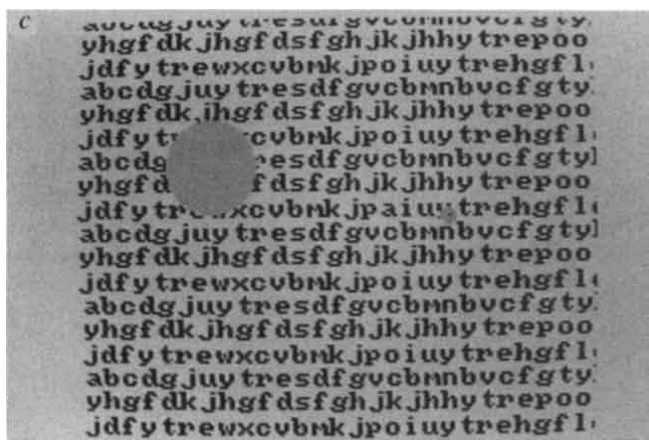
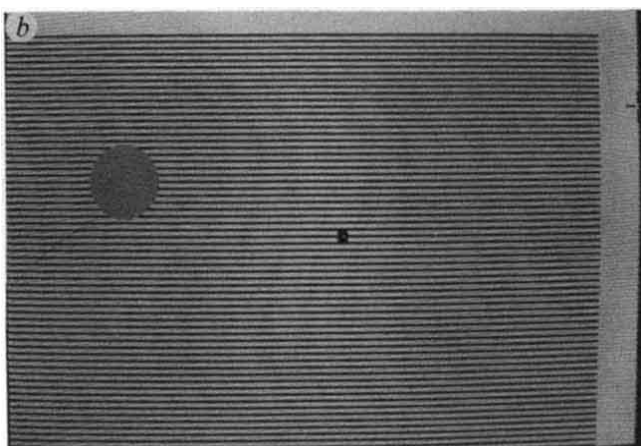
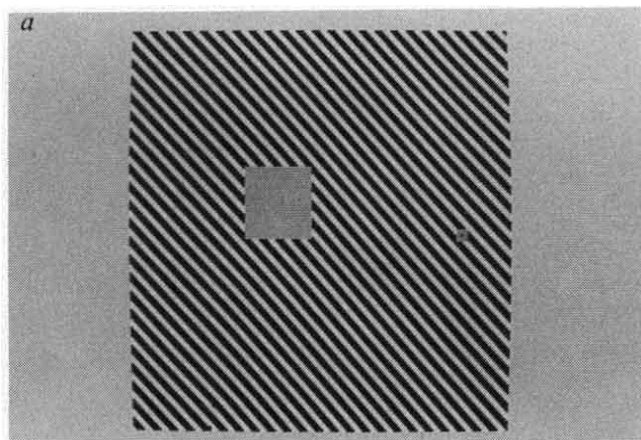


FIG. 3a, In this display the surround consisted of a counterphase flickering grating instead of twinkling noise. Fading and filling in took longer but could still be seen. b, In this display we have stationary continuously visible horizontal lines instead of a counterphase flickering grating. Fading and filling in can be seen after 15–20 s of steady eccentric fixation. We find that one out of every five or six subjects cannot achieve the steady fixation required for this effect. Also, on some trials the entire pattern fades (as in 'Troxler fading'³⁰) instead of the square, an effect that is rarely seen in Fig. 1. c, The square is displayed against a 'nonsense' matrix of English letters. Again, the square fades and gets filled in by the alphabets, but subjects could not actually read the letters in the filled-in region.

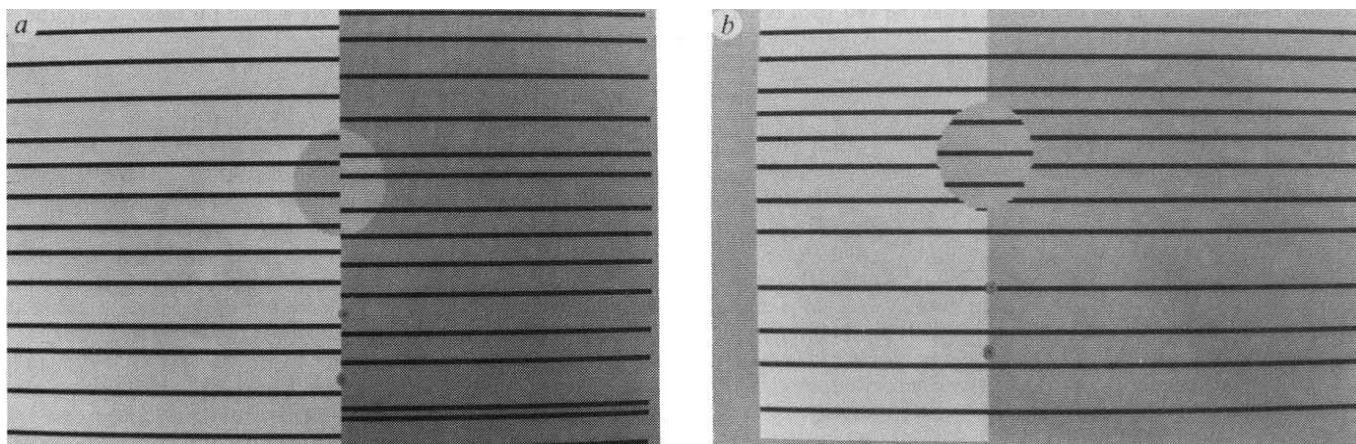


FIG. 4 Stimulus configuration used to demonstrate the effect of illusory contours on the filling-in process. *a*, the background was pink on the left (depicted light grey here) and green (depicted dark grey here) on the right and the circular disk was an equiluminous grey. Filling in occurred from both sides and formed a 'border' between the two filled-in colours that coincided

with the illusory contour. In the absence of the central vertical illusory contour (*b*) the border appeared indistinct and unstable. The filling in of the disk also took longer in *b*, perhaps because of the illusory contours associated with the chromatic borders. Eleven naive subjects reported this effect; filling in took place earlier in *a* than in *b* in 195 out of 220 trials.

introduced that coincided with the chromatic border and also continued across the centre of the grey disk (Fig. 4*a*). On steady eccentric fixation, subjects now reported that the colours filled in the disks from the two sides to form an illusory colour border corresponding to the illusory contour. If the illusory contour was not present the colour border within the scotoma usually appeared 'washed out', indistinct and unstable. (If the illusory contours coincided with the borders of the disk itself (Fig. 4*b*) the fading or filling-in processes were delayed.) We may conclude from this that the filling-in process is strongly influenced by early image segmentation, produced by illusory contours.

Finally, we wondered whether the filling in of the natural blind spot is also an active process that involves spatial integration. To explore this we used a 10°-wide red annulus (ring) whose inner margins just overlapped the outer margins of the blind spot, viewed monocularly. The red colour filled in the blind spot so that the stimulus looked like a homogeneous red disk rather than a ring. Yet when we introduced several concentric thin red rings around the annulus the blind spot appeared filled with rings instead of a homogeneous red, an effect that was confirmed by four naive subjects. Thus what fills in the

blind spot is not merely what is immediately around it but also the pattern from the remote surround.

Considerable progress has recently been made in understanding the nature of neural representation underlying perception²³⁻²⁹. Our results suggest that the rules used by the visual system to fill in perceptual scotomas may help to elucidate further some of these processes. One intriguing possibility is that filling in simply reflects the activity of higher extrastriate neurons whose large receptive fields may straddle the boundaries of the grey square. Once the neurons that signal the square's border become fatigued, the signals from the surround may get misattributed to the region corresponding to the scotoma. A second possibility is that feedback pathways^{11,24} from extrastriate areas are used to create a representation of the surround in the portion of area 17 that corresponds to the scotoma. This theory complements, rather than negates, Sir David Brewster's hypothesis⁶ that the filling in is achieved by 'the Divine Artificer', and it has the additional merit of being more easily testable.

Note added in proof. In patients with small (1°) retinal scotomas we find that filling in occurs for homogeneous colours and fine-grained static noise but not for dynamic twinkling noise (Fig. 1). □

Received 27 December 1990; accepted 27 February 1991.

- Bender, M. B. & Teuber, H. L. *Arch. Neur. Psychiat.* **55**, 627-658 (1946).
- Sergent, J. *Brain* **3**, 347-373 (1988).
- Weizsaecker, I. L. in *Current Problems in Animal Behavior* (eds W. H. Thorpe and O. Zangwill), 30-58 (Cambridge University Press, Cambridge, 1961).
- Poppel, E. *Nature* **320**, 523-525 (1986).
- Teuber, H. L. in *Brain and Conscious Experience* (ed. J. C. Eccles) 182-216 (Springer, New York, 1966).
- Brewster, D. *Letters on Natural Magic*, addressed to Sir Walter Scott, Bart. (John Murray, London, 1832).
- Ramachandran, V. S. & Anstis, S. M. 11th European Conference on Visual Perception. Bristol, UK, 1987.
- Ramachandran, V. S. *Perception* **19**, 273 (1990); *The Artful Brain* (Oxford University Press, in the press).
- Zeki, S. M. *Nature* **335**, 311-317 (1988).
- Allman, J., Meizins, F. & McGuinness, E. L. *Perception* **14**, 105-126 (1985).
- Van Essen, D. C. *A. Rev. Neurosci.* **2**, 227-263 (1979).
- Livingstone, M. S. & Hubel, D. H. *J. Neurosci.* **7**, 3416-3468 (1987).
- Krauskopf, J. *Am. J. Psychol.* **80**, 632-637 (1961).
- Crane, H. D. & Piantanida, T. P. *Science* 1078-1079 (1983).
- Yarbus, A. L. *Eye Movement and Vision* (Engl. Trans. Haigh, B.; ed. Riggs, L. A.) (Plenum, New York, 1967).

- Kanizsa, G. *Organization in Vision* (Prager, New York, 1979).
- Gregory, R. L. *Nature* **238**, 51-52 (1972).
- Von Der Heydt, R. E., Peterhans, E. & Baumgartner, G. *Science* **224**, 1260-1262 (1978).
- Ramachandran, V. S. *Perception* **14**, 127-134 (1985).
- Ramachandran, V. S. *Soc. Neurosci. Abstr.* **15**, 624.
- Ramachandran, V. S. *Perception and Psychophysics* **39**, 361-373 (1986).
- Ramachandran, V. S. *Nature* **328**, 645-647 (1987).
- Ramachandran, V. S. in *AI and the Eye* (eds Blake, A. & Troscianko, T.) 21-77 (Wiley, Bristol, 1990).
- Finkel, L. H. & Edelman, G. M. *J. Neurosci.* **9**, 3188-3208 (1989).
- Grossberg, S. & Mingolla, E. *Psychol. Rev.* **92**, 173-211 (1985).
- Shepard, R. in *Perceptual Organization* (eds Kubovy, M. & Pomerantz, J. R.) 279-343 (Erlbaum, New Jersey, USA, 1981).
- Julesz, B. *Foundations of Cyclopean Perception* (University of Chicago Press, Chicago, 1971).
- Poggio, T., Torre, V. & Koch, C. *Nature* **138**, 645-647 (1985).
- Rock, I. *The Logic of Perception* (MIT, Cambridge, Massachusetts, 1983).
- Troxler, D. in *Ophthalmologische Bibliothek* **II**, PT2, 1-53 (1804).

ACKNOWLEDGEMENTS. We thank W. Aiken, A. Saluda, D. Rogers-Ramachandran, F. Crick, T. Sejnowsky, P. Churchland and C. Ramachandran for stimulating discussions, and the Airforce Office of Scientific Research for support. R.L.G. was a McDonnell-Pew visiting scholar at UCSD.

Peptide selection by MHC class I molecules

Ton N. M. Schumacher*, Marloes L. H. De Bruijn†, Leen N. Vernie*, W. Martin Kast†, Cornelis J. M. Melief†, Jacques J. Neefjes* & Hidde L. Ploegh*

*Departments of *Cellular Biochemistry and †Immunology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands

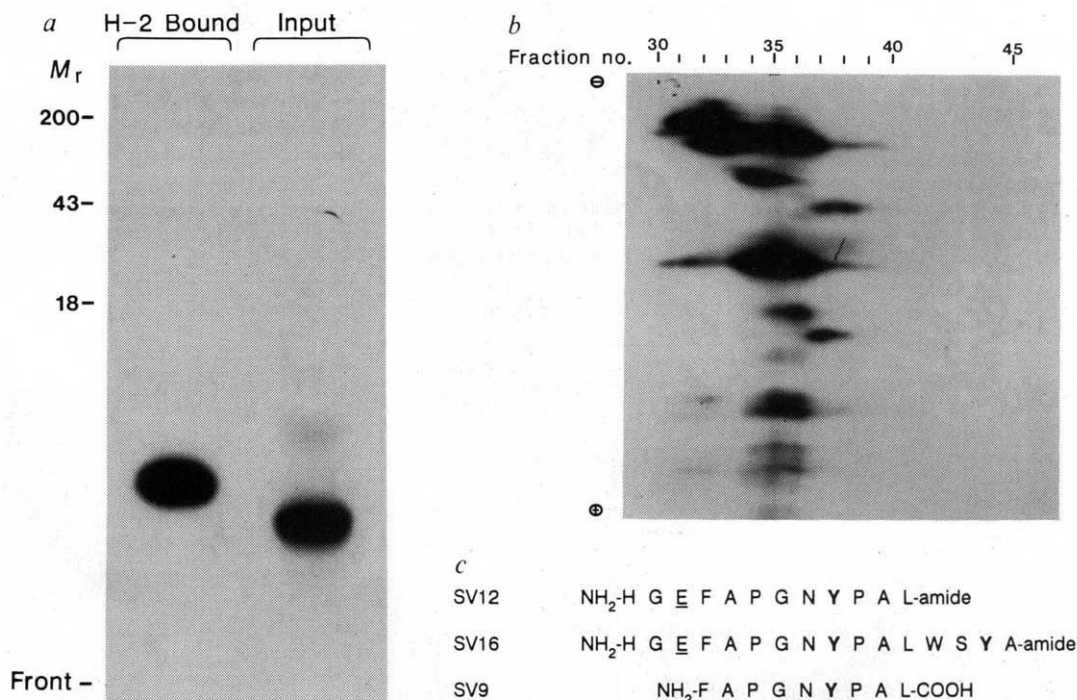
SYNTHETIC peptides have been used to sensitize target cells and thereby screen for epitopes recognized by T cells¹⁻⁷. Most epitopes of cytotoxic T lymphocytes can be mimicked by synthetic peptides of 12-15 amino acids⁸. Although in specific cases, truncations of peptides improves sensitization of target cells^{8,9}, no optimum length for binding to major histocompatibility complex (MHC) class I molecules has been defined. We have now analysed synthetic peptide captured by empty MHC class I molecules of the mutant cell line RMA-S. We found that class I molecules preferentially bound

short peptides (nine amino acids) and selectively bound these peptides even when they were a minor component in a mixture of longer peptides. These results may help to explain the difference in size restriction of T-cell epitopes between experiments with synthetic peptides and those with naturally processed peptides¹⁰⁻¹².

MHC class I molecules in the murine cell line RMA-S are unable to engage endogenous peptides¹³⁻¹⁶. Empty class I molecules formed in this cell line are unstable at 37 °C, but are expressed in considerable quantities at the surface at reduced temperatures (optimum, 26 °C; notation for this cell line at 26 °C, RMA-S(26)) and can be used to measure binding of peptide to class I molecules^{14,15}.

A peptide preparation derived from Sendai virus nucleoprotein and identified as a cytotoxic T-lymphocyte (CTL) epitope (SV12 peptide; residues 321-332 of the nucleoprotein: HGEFAPGNYPAL in the single-letter amino-acid code) binds and stabilizes empty H-2 K^b and D^b molecules¹⁷. A 16-residue Sendai peptide extended at its carboxy terminus (SV16 peptide; residues 321-336; sequences in Fig. 1c) could also stabilize empty class I molecules and sensitize targets for lysis by CTL. SV12 was radiolabelled by iodination, presented to RMA-S(26) cells, and class I-associated peptide was analysed on an SDS-tricine gel system¹⁸. The class I-bound material did not comigrate

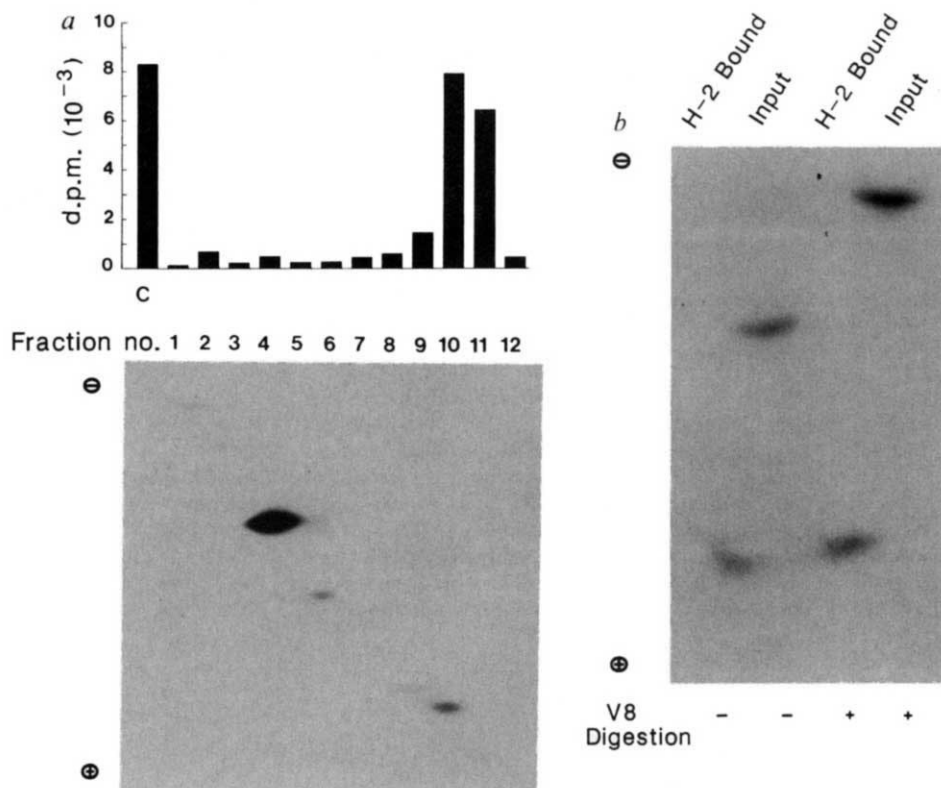
FIG. 1 *a*, MHC class I-bound peptide is not identical to the input peptide. RMA-S cells were cultured for 48 h at 26 °C, washed and incubated for 1 h with 5 μ M [¹²⁵I]-labelled Sendai nucleoprotein peptide (residues 321-332, SV12 peptide) at 26 °C. After washing and lysis, class I molecules were isolated by immunoprecipitation and analysed by electrophoresis on a SDS-tricine gel. Relative molecular mass markers are indicated on the left (M_r). For small peptides, separation is based both on charge and SDS-binding properties, rather than the true molecular mass of the peptide. Note that the peptide bound to H-2 molecules does not comigrate with the major peptide in the input material. *b*, Heterogeneity of SV12 as revealed by two-dimensional HPLC and isoelectric focusing analysis. SV12 peptide was labelled by iodination and separated in the first dimension by reversed-phase HPLC. HPLC fractions were lyophilized and resuspended in RPMI. Small aliquots of all fractions were analysed by isoelectric focusing²⁰. Only the region of the gel containing the majority of the labelled peptide is shown. Note that overexposure of one-dimensional isoelectric-focusing analysis of the different HPLC fractions reveals the presence of minor contaminants that elute in the vicinity of the main peak. *c*, Sendai nucleoprotein-derived peptides that can sensitize target cells for lysis by H-2K^b-restricted CTLs. Sequences are given in single-letter code. V8 protease recognition sites are underlined. Preferred sites of iodination catalysed by chloramine T (Y) are in bold. Amino-acid composition was determined on hydrolysed SV12 peptide. Yield of the separate amino acids relative to the most C-terminal amino acid (leucine) is given. Expected values are indicated between brackets (moles per mole of peptide). His, 0.7 (1); Gly, 1.6 (2); Glu, 1.0 (1); Phe, 1.0 (1); Ala, 2.0 (2); Pro, 1.7 (2); Asp, 0.7 (1); Tyr, 0.8 (1); Leu, 1.0 (1). Sequence analysis performed on an ABS gas phase sequencer shows that the peptide is of the expected sequence. The synthesis strategy probably does not result in peptides with internal deletions, because growing peptide chains that fail to couple are terminated with a blocking agent and cannot engage in further chain elongation. Truncations arise as a consequence of premature termination in the course of synthesis (N-terminal truncations), or partial hydrolysis during cleavage of



peptide of the resin (N- or C-terminal truncations). Further possible contaminants result from incomplete deprotection and deamidation.

METHODS. Peptides were synthesized by standard t-boc chemistry on a Biosearch SAM 2 peptide synthesizer, dissolved in PBS medium and stored at 4 °C. Peptides were iodinated in a reaction catalysed by chloramine T (ref. 21) to a specific activity of about 50 Ci mmol⁻¹. The mouse lymphoma cell lines transformed with Rauscher virus, RMA and RMA-S (refs 22, 23), were maintained in DMEM medium supplemented with 10% FCS and antibiotics. Before use, cells were cultured for 48 h in DMEM-10% FCS supplemented with 20 mM HEPES buffer and antibiotics (10⁶ cells per ml) in a waterbath at 26 °C. Cells were then incubated for 1 h with 5 μ M [¹²⁵I]-labelled Sendai nucleoprotein peptide at 26 °C. Cells were washed three times with PBS and lysed on ice in TX-100 lysis buffer. Class I molecules were immunoprecipitated and analysed by electrophoresis on SDS-tricine gel¹⁸. Gels were then frozen and exposed to Kodak XAR-5 films at -70 °C. Peptide binding assays were performed in the absence of FCS to avoid peptide modifications by serum components⁷. *b*, Radiolabelled SV12 was resolved by HPLC on a Beckman ODS C-18 RP 4.6 $\times$ 250-mm reversed-phase column. After initial isocratic elution in 0.1% TFA for 2 min, linear gradient elution to 0.1% TFA-50% acetonitrile was done for 50 min. Fractions (1 ml) were collected, lyophilized and run on one-dimensional isoelectric focusing gels²⁴.

FIG. 2 *a*, Class I binding peptide can be isolated by preparative isoelectric focusing. SV12 peptide was radiolabelled by iodination and subsequently separated by preparative one-dimensional isoelectric focusing done as described²⁴, but without urea in the gel. After isoelectric focusing, peptide was eluted from 1-cm gel slices, lyophilized and resuspended in RPMI. Small aliquots of the different fractions were analysed by rerun on one-dimensional isoelectric focusing (bottom). The fractions were used in a peptide-binding assay (top), as described in Fig. 1. RMA-S(26) cells were incubated with a 1:10 dilution in DMEM-Hepes of fractions 1–12. After washing and lysis of the cells, class I antigens were immunoprecipitated, and the presence of class I-associated peptide was determined by gamma-spectrometry. The binding of control, unfractionated ¹²⁵I-labelled SV12 peptide to class I molecules is shown on the left of the figure. *b*, V8 protease insensitivity of class I captured peptide. RMA-S(26) cells were incubated with ¹²⁵I-labelled SV12 peptide as described in Fig. 1. Class I-associated peptide was isolated by immunoprecipitation of class I antigens. Equal amounts of d.p.m. of the class I-associated peptide and the input SV12 peptide were incubated for 5 min at 95 °C in 200 mM Tris, 200 mM tricine, 0.2% SDS pH 8.25, cooled on ice and subsequently incubated for 30 min at 37 °C with or without 0.5 mg per ml V8 protease (Sigma), which cleaves the C terminal from glutamic



acid. Peptides were then analysed by one-dimensional isoelectric focusing²⁴ as described, but without urea.

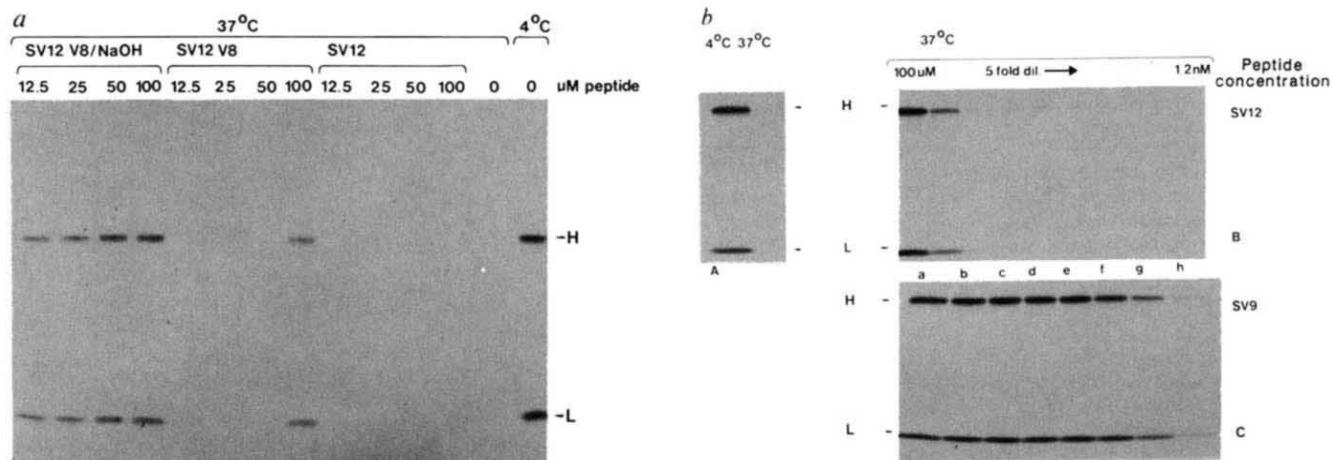


FIG. 3 *a*, A combined V8-NaOH treatment of the SV12 peptide improves the ability to stabilize empty class I molecules. The SV12 peptide was digested with V8 protease or mock digested as described in Fig. 2*b*. Subsequently, peptides were either diluted 20-fold in 50 mM NaOH, or 40-fold in 25 mM NaOH, 500 mM Tris, pH 7.8, and incubated at 95 °C for 30 min. Peptides diluted in 50 mM NaOH were diluted twofold in 1 M Tris, pH 7.8, afterwards. One-dimensional isoelectric focusing confirmed that incubation at 95 °C at neutral pH does not lead to deamidation, whereas incubation in 50 mM NaOH leads to considerable deamidation. Mock-treated, V8-treated and V8-NaOH-treated SV12 peptides were used in a class I stabilization assay. RMA-S(26) cells were cell-surface labelled by lactoperoxidase catalysed iodination²⁰. After washing and lysis, cell lysates were incubated with the different peptides for 1 h at the concentrations indicated. Lysates were then exposed for 1 h to 4 °C or 37 °C, class I antigens were isolated using an anti H-2^b serum and immunoprecipitates were analysed on a 12% SDS-PAGE gel. The V8-NaOH-treated SV12 peptide stabilizes class I molecules at much lower concentrations than the original SV12 peptide preparation. *b*, Stabilization of empty class I molecules by SV12 and SV9 peptides. RMA-S(26) cells were cell-surface labelled by lactoperoxidase-

catalysed iodination. After washing and lysis, cell lysates were incubated with the different peptides for 1 h at the concentrations indicated in Methods. Lysates were then exposed for 1 h to 4 °C or 37 °C. Class I antigens were isolated using an anti H-2^b serum and immunoprecipitates were analysed on a 12% SDS-PAGE gel. Note that the SV9 peptide can stabilize class I molecules when present at concentrations as low as 6 nM, whereas the SV12 peptide is only active at concentrations >20 μM. On the assumption that only SV9 is capable of stabilization this would imply that the amount of contaminating 'SV9 peptide' in this preparation is less than 0.1%. METHODS. Cells were cell-surface labelled at 0 °C using lactoperoxidase-catalysed iodination as described²⁴, washed thrice with PBS and lysed on ice in TX-100 lysis buffer (10 mM Tris, pH 7.8, 140 mM NaCl, 1% Triton X-100, 1 mM PMSF, 1 μg ml⁻¹ trypsin inhibitor, 30 mTIU (trypsin inhibitory units) ml⁻¹ aprotinin). Lysates were incubated with peptides for 1 h at 4 °C, in *a* at the concentrations indicated, and in *b* at the following concentrations: a, 100 μM; b, 20 μM; c, 4 μM; d, 800 nM; e, 160 nM; f, 32 nM; g, 6.4 nM; h, 1.2 nM. Lysates were then exposed for 1 h to either 4 °C or 37 °C. Class I antigens were immunoprecipitated using a rabbit serum raised against purified H-2^b.

with the major or any obvious minor peptide in the input material (Fig. 1a). A peptide of similar mobility was selected by RMA-S(26) from radiolabelled SV16 (Fig. 4A). Contaminating peptides are likely to include truncations (see legend to Fig. 1). For the two synthetic peptides containing the same core sequence, a single truncated peptide might therefore be able to associate with class I molecules. Alternatively, SV12 and SV16 could be trimmed by proteolysis to the same length after binding to class I molecules. To check for the presence of possible peptide

contaminants, SV12 was radiolabelled, and separated by two-dimensional HPLC–isoelectric focusing. Several minor contaminants were present (Fig. 1b).

SV12 preparation was separated by isoelectric focusing (Fig. 2a, bottom), and the resulting fractions were used in peptide binding assays. No class I-binding peptide could be detected in the peak fraction (fraction 4), and only fractions 10 and 11, which contain a small part of the input material, contained class I-binding peptide (Fig. 2a, top). Class I-bound peptide and the material in fraction 10 were indistinguishable in size (tricine gel) and charge (isoelectric focusing: not shown). These results suggest that a proteolytic conversion, catalysed either by the class I molecules themselves, or by associated proteases, is not responsible for the generation of the peptide. The material in isoelectric-focusing fraction 10 and the class I-eluted material were subjected to Edman degradation. In both cases the label was recovered after six cycles of degradation, suggesting that the three N-terminal residues of SV12 were absent from this peptide. Indeed, the class I-captured peptide was insensitive to treatment with V8 protease (Fig. 2b).

The removal of the first three residues of SV12 results in a peptide with a higher isoelectric point than that of the H-2-binding peptide. The C-terminal leucine-amide is therefore probably lost or modified. Because this leucine can be replaced by only a few amino acids in SV16 and still result in target-cell sensitization (W.M.K. *et al.*, unpublished observation), deletion seems unlikely.

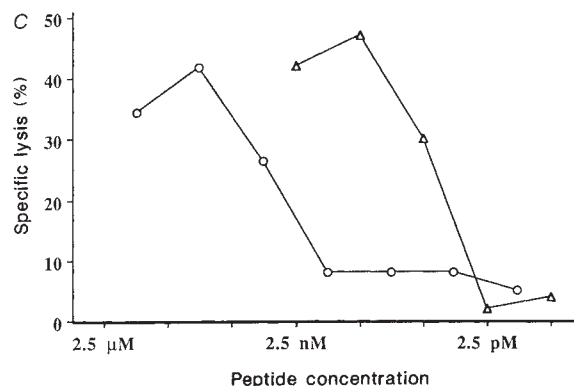
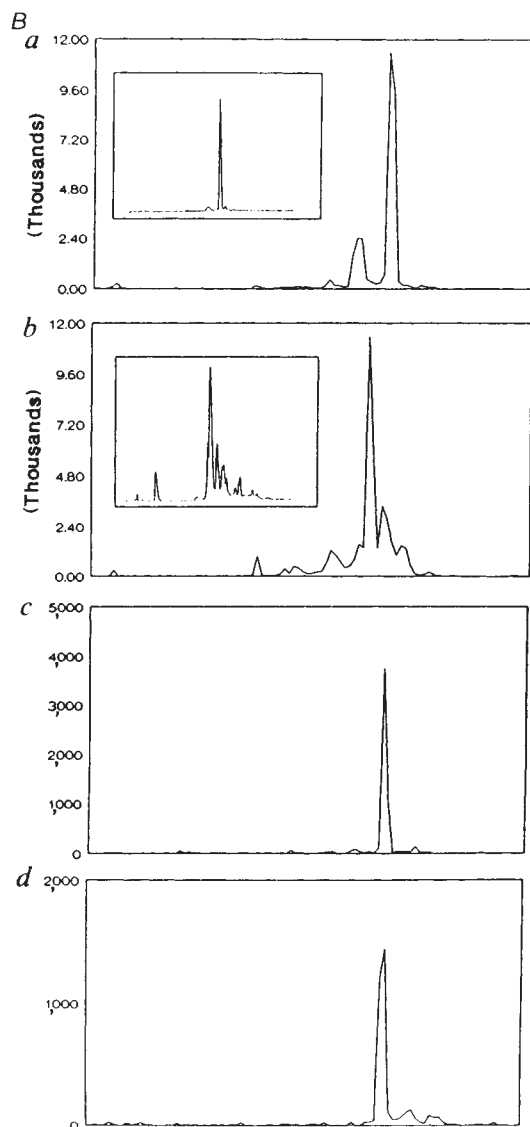
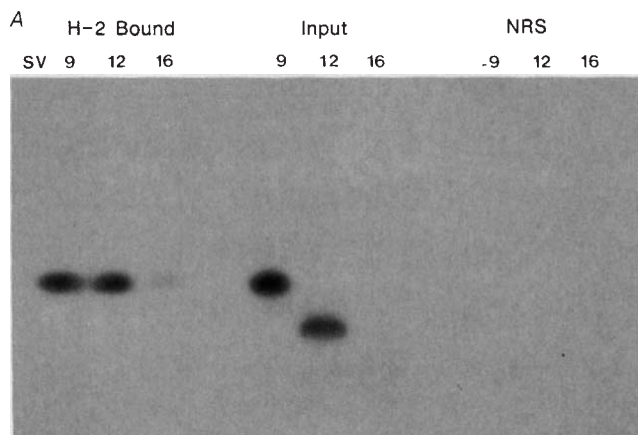


FIG. 4 A, Analysis of input and MHC-captured SV peptides on SDS–tricine gel. SV9, SV12, and SV16 peptides were labelled by chloramine T-catalysed iodination and offered to RMA-S(26) cells as described in the legend to Fig. 1. MHC captured and input peptides were analysed on SDS–tricine gel. H-2: H-2 captured peptide; Input: input peptide; NRS: normal rabbit serum control. Note comigration of MHC-bound peptides for SV9, SV12 and SV16. B, HPLC analysis of input and MHC-captured SV peptides. a, Unlabelled (inset) and labelled SV12 peptide; b, unlabelled (inset) and labelled SV9 peptide; c, MHC-captured SV12 peptide; d, MHC-captured SV9 peptide. C, Target cell sensitization by SV16 and SV9 peptides. SV16 (○) and SV9 (△) peptides were tested at different concentrations for sensitization of target cells to lysis by a Sendai virus-specific CTL clone^{17,26}. Half-optimum specific ⁵¹Cr release was reached at a concentration of 10 nM SV16 and 20 pM SV9 peptide.

METHODS. B, H-2-captured and input-radiolabelled peptides were prepared as described in the legend to Fig. 1. Samples were dissolved in 0.05% TFA and analysed by reverse phase HPLC on a C18 reverse phase HPLC column (Millipore Waters). Buffers: A, H₂O–0.05% TFA; B, acetonitrile–0.05% TFA. Elution profile 0–3 min 95→95% A, 3–5 min 95→85% A, 5–23 min 85→60% A, 23–34 min 60→35% A. Fractions (0.3 min) were collected and radioactivity in the individual fractions was determined by gamma-spectrometry. The peak of radioactivity was in fraction 61 for b, and in fractions 64–65 for a, c and d. Material captured from SV16 likewise eluted at fractions 64–65 (not shown). C, B6 mouse embryo cells were cultured for 2 days in the presence of 50 U per ml recombinant interferon- γ to increase class I expression. Cloned CTL (2×10^4) were added to 2×10^3 ⁵¹Cr-labelled B6 MEC in a final volume of 200 μ l in 96-well U-bottomed plates. Cells were incubated for 6 h at 37 °C in humidified air with 5% CO₂. ⁵¹Cr release was calculated as described²⁶. Peptides were present during the whole assay at the concentrations indicated.

Empty class I molecules in detergent lysates rapidly change their conformation at 37 °C, which can be prevented by incubation with presentable peptides¹⁵. SV12 digested with V8 protease, and V8-digested SV12 partially deamidated by dilute NaOH, were tested for their ability to stabilize empty class I molecules. The combination of V8 digestion and deamidation yields a peptide preparation much more potent at stabilizing empty class I molecules (Fig. 3a). SV9 peptide with a free carboxyl group binds to and stabilizes class I molecules in detergent lysates at concentrations as low as 6 nM (>4,000-fold lower concentration than required for stabilization by the SV12 peptide; Fig. 3b). If a peptide corresponding to SV9 were solely responsible for the stabilizing effect of SV12, then it probably constitutes less than 0.1 per cent of total peptide. The material captured from SV9, SV12 and SV16 comigrates with input SV9 on SDS-tricine gel (Fig. 4A), one-dimensional isoelectric focusing (not shown) and HPLC (shown for SV12 in Fig. 4B) under conditions where SV9, SV12 and SV16 are resolved. The ability of the different Sendai virus peptides to sensitize target cells for lysis by Sendai-specific CTL shows that SV9 can induce sensitization at 100–1,000-fold lower concentrations than the longer peptides derived from Sendai virus (Fig. 4C). Concentrations of the SV9 peptide in the picomolar range are sufficient to sensitize for lysis by CTL.

Both ends of the groove in HLA-A2 and A68 seem to be closed¹⁹ and could therefore put an upper limit on the size of binding peptides. Peptides longer than nine amino acids might not fit into the groove as tightly as those of optimal size (8 or 9 amino acids), but they could bind at one end of the groove, with one of their sides projecting away from it. Alternatively, both ends of the peptide could be neatly tucked in, the middle part interacting less tightly. A free terminal carboxyl group is apparently essential for high-affinity binding of the Sendai peptide to class I molecules. Therefore, the size limit for peptides might not only reflect the boundaries of the peptide-binding groove, but also the positioning of the free carboxyl group. That only few amino-acid substitutions are allowed for the SV9 core in SV16, whereas residues outside this core can be replaced by all other amino acids (W.M.K. *et al.*, unpublished observation), suggests that there might also be a lower limit to the size of class I binding peptides. The set of endogenous peptides extracted with trifluoroacetic acid from intact cells and capable of sensitizing for lysis by CTL is determined by the class I type of the donor cell¹⁰, although direct selection of peptide by class I molecules themselves was not shown. These endogenously derived peptides are shorter and more potent than the T-cell epitopes deduced from experiments using synthetic peptides, and are similar in size to the Sendai peptide. Endogenous peptides directly isolated from class I molecules in cells infected with vesicular stomatitis virus were also remarkably homogeneous and composed of only eight residues¹¹. This homogeneity has been suggested to be imposed by the peptide-generating or translocating machinery, or to be an intrinsic property of class I molecules^{10–12}. Our experiments show that class I molecules themselves are remarkably selective in peptide binding. □

18. Schagger, H. & von Jagow, G. *Analyt. Biochem.* **166**, 368–379 (1987).
19. Bjorkman, P. J. & Parham, P. A. *Rev. Biochem.* **59**, 253–288 (1990).
20. Phillips, D. R. & Morrison, M. *Biochem. biophys. Res. Commun.* **40**, 284–289 (1970).
21. Hunter, W. M. & Greenwood, F. C. *Nature* **194**, 495–496 (1962).
22. Ljunggren, H. G. & Kärre, K. *J. exp. Med.* **162**, 1745–1759 (1985).
23. Kärre, K., Ljunggren, H. G., Piontek, G. & Kiesling, R. *Nature* **319**, 675–678 (1986).
24. Neefjes, J. J., Breur-Vriesendorp, B. S., Van Severen, G. A., Ivanyi, P. & Ploegh, H. L. *Hum. Immun.* **16**, 169–181 (1986).
25. Hämmerling, G. J., Hämmerling, U. & Lemke, H. *Immunogenetics* **8**, 433–445 (1979).
26. Kast, W. M., Bronkhorst, A. M., de Waal, L. P. & Melief, C. J. M. *J. exp. Med.* **164**, 723–738 (1986).

ACKNOWLEDGEMENTS. We thank E. Palmieri and S. Nathenson for the gift of rabbit anti H-2^b serum, J. Nieland and R. Brandt for technical assistance, R. van der Valk, P. van Balen and T. Lutz for assistance with HPLC and peptide iodinations, D. Wiley for helpful comments, and W. J. Weijer and H. J. Bak from Eurosequence, Groningen, for amino-acid and sequence analysis. This research was supported by the Netherlands Cancer Foundation and the Netherlands Organisation for Scientific Research.

Enhancement of T-cell activation by the CD43 molecule whose expression is defective in Wiskott–Aldrich syndrome

John K. Park*, Yvonne J. Rosenstein*†, Eileen Remold-O'Donnell§||, Barbara E. Bierer*‡, Fred S. Rosen§||, & Steven J. Burakoff*§

* Division of Pediatric Oncology, Dana Farber Cancer Institute,

† Division of Hematology, Brigham and Women's Hospital,

|| The Center for Blood Research and the Departments of § Pediatrics and

‡ Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA

CD43 (sialophorin, leukosialin, leukocyte large sialoglycoprotein), a heavily sialylated molecule found on most leukocytes and platelets, was initially identified as a major glycoprotein of mouse, rat and human T cells^{1–8}. CD43 expression is defective on the T cells of males with the Wiskott–Aldrich syndrome, an X chromosome-linked recessive immunodeficiency disorder⁹. Affected males are susceptible to opportunistic infections and do not respond to polysaccharide antigens, reflecting defects in cytotoxic and helper T-cell functions. Anti-CD43 monoclonal antibodies have a modest costimulatory effect on T cells, natural killer cells, B cells and monocytes^{10–14}, and one such antibody has been shown to activate T cells directly¹⁵. To investigate a possible physiological role for CD43, a complementary DNA encoding the human protein^{16,17} was introduced into an antigen-responsive murine T-cell hybridoma¹⁸. We observed that CD43 enhances the antigen-specific activation of T cells and that the intracellular domain of CD43, which is hyperphosphorylated during T-cell activation^{19–21}, is required for this function. We also found that antigen-presenting cells can bind specifically to immobilized purified CD43 and that the binding can be inhibited by liposomes containing CD43 as well as by anti-CD43 monoclonal antibodies.

The murine T-cell hybridoma By155.16 produces interleukin-2 (IL-2) following stimulation of its T-cell receptor/CD3 complex (TCR-CD3) by anti-CD3 monoclonal antibody. The TCR of By155.16 is specific for HLA-DR antigens such as those expressed by Daudi, a Burkitt's lymphoma-derived B-cell line. As a result, By155.16 can be stimulated to produce IL-2 by coculture with Daudi. But the particular subclone of By155.16 studied here can only be stimulated to produce IL-2 if it has been previously transfected with genes encoding human T-cell accessory molecules such as CD4, which recognizes HLA-DR, or CD2, which recognizes LFA-3^{18,22,23}. The dependence of the Daudi-induced activation of this subclone on such accessory molecules was exploited to determine whether CD43 could similarly function as an accessory or costimulatory molecule.

Received 13 December 1990; accepted 1 March 1991.

1. Townsend, A. R. M. *et al.* *Cell* **44**, 959–968 (1986).
2. Maryanski, J. L., Pala, P., Corradin, G., Jordan, B. R. & Cerottini, J. C. *Nature* **324**, 578–579 (1986).
3. Gotch, F., Rothbard, J., Howland, K., Townsend, A. & McMichael, A. *Nature* **326**, 881–882 (1987).
4. Kast, W. M. *et al.* *Cell* **59**, 603–614 (1989).
5. Oldstone, M. B. A., Whitton, J. L., Lewicki, H. & Tishon, A. *J. exp. Med.* **168**, 559–570 (1988).
6. Gotch, F., McMichael, A. & Rothbard, J. *J. exp. Med.* **166**, 2057 (1988).
7. Moore, M. W., Carbone, F. R. & Bevan, M. J. *Cell* **54**, 777–785 (1988).
8. Townsend, A. & Bodmer, H. A. *Rev. Immun.* **7**, 601–624 (1989).
9. Reddehase, J. M., Rothbard, J. B. & Koszinowski, U. H. *Nature* **337**, 651–653 (1989).
10. Rötzschke, O. *et al.* *Nature* **348**, 252–254 (1990).
11. Van Bleek, G. M. & Nathenson, S. G. *Nature* **348**, 213–216 (1990).
12. Elliot, T., Townsend, A. & Cerundolo, V. *Nature* **348**, 195–197 (1990).
13. Townsend, A. *et al.* *Nature* **340**, 443–448 (1989).
14. Ljunggren, H. G. *et al.* *Nature* **346**, 476–480 (1990).
15. Schumacher, T. N. M. *et al.* *Cell* **62**, 563–567 (1990).
16. Townsend, A. *et al.* *Cell* **62**, 285–295 (1990).
17. Kast, W. M. *et al.* *Proc. natn. Acad. Sci. U.S.A.* (in the press).

† Permanent address: Instituto de Investigaciones Biomedicas, UNAM, Mexico.

A cDNA encoding either the wild-type CD43 protein (CD43) or a CD43 protein with the intracellular domain deleted (Δ CD43) was inserted into the plasmid pFneo²⁴, which contains the G418 resistance gene (*neo^r*) (Fig. 1). By155.16 cells were transfected by electroporation²⁵ with pFneo, pCD43Fneo, or p Δ CD43Fneo, selected for resistance to G418, and subcloned by limiting dilution. From each transfection group, clones expressing equivalent levels of TCR-CD3 and/or the extracellular domain of CD43 where appropriate were selected for further study (Fig. 2a and b). Also selected were CD43⁻, CD43⁺ and Δ CD43⁺ clones which were TCR-CD3⁻. To assess the ability of the different clones to generate IL-2, they were stimulated with anti-CD3 monoclonal antibody. Those clones expressing TCR-CD3 responded equivalently (Fig. 3a), whereas those expressing equivalent levels of CD43 or Δ CD43 but no detectable TCR-CD3 produced <20 U IL-2. To determine whether CD43 could have a costimulatory role in T-cell activation induced by antigen-presenting cells, the clones were cocultured with Daudi. Of the TCR-CD3⁺ clones, the CD43⁺ clone responded to Daudi stimulation, but the CD43⁻ and Δ CD43⁺ clones did not (Fig. 3a). Likewise, none of the TCR-CD3⁻ clones responded to Daudi stimulation (IL-2 production <20 units). In the presence of TCR-CD3, therefore, CD43 can serve as an

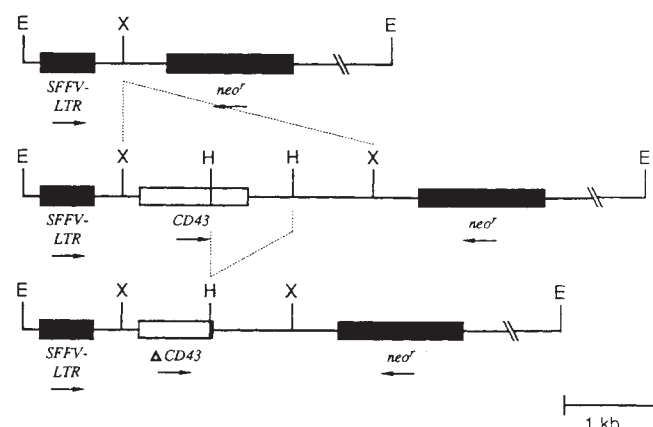


FIG. 1 Plasmid constructs. A full length CD43 cDNA clone (gift of C. S. Shelley and Michael Carroll) was subcloned into the pFneo plasmid (top; gift of R. Germain) to generate the pCD43Fneo plasmid (middle). Plasmid pCD43Fneo was in turn modified to generate the p Δ CD43Fneo plasmid (bottom). The dotted lines between the three plasmids indicate areas of insertions and deletions. Plasmid pFneo and its derivatives contain the G418 resistance gene *neo^r* and the spleen focus-forming virus long terminal repeat (SFFV-LTR) which promotes the expression of 5' $\rightarrow$ 3'-oriented genes located near its 3' terminus. Plasmid pCD43Fneo-transfected cells can express the full-length (381 amino acids) form of CD43, and p Δ CD43Fneo-transfected cells a truncated form of CD43 retaining all the 259 predicted extracellular and transmembrane domain amino acids, but only the first of the 123 predicted intracellular domain residues. E, *EcoRI* site; X, *XbaI* site; H, *HaeIII* site; A, *Apal* site; Arrows indicate direction of transcription.

METHODS. The CD43 cDNA clone HPB2.5 was isolated from an HPB-ALL cDNA library and subcloned into the *NotI* and *HindIII* sites of π H3M to generate π H3M-CD43. A 2.8-kb *XbaI* fragment containing the CD43 cDNA was excised from π H3M-CD43 and ligated into the *XbaI* site of pFneo. An 8.4-kb clone containing the CD43 cDNA in the proper orientation with respect to the SFFV-LTR (pCD43Fneo) was isolated for further manipulation. To generate p Δ CD43Fneo, pCD43Fneo was digested with *EcoRI* and *Apal* to yield: 2.1-kb *EcoRI*-*Apal*, 1.2-kb *Apal*, and 5.1-kb *Apal*-*EcoRI* fragments. The 2.1-kb and 1.2-kb fragments were further digested with *HaeIII* and the resulting 2.1-kb *EcoRI*-*HaeIII* and 0.9-kb *HaeIII*-*Apal* fragments were purified and religated to the 5.1-kb *Apal*-*EcoRI* fragment. The net result was the deletion of a 0.3-kb *HaeIII* fragment encoding for 122 of the 123 intracellular amino acids. (The presence of multiple *HaeIII* sites within the 5.1-kb *Apal*-*EcoRI* fragment necessitated this approach.) Because this deletion also removed the stop codon of the original peptide-coding sequence, a new stop codon beginning 49 base pairs 3' to the ligation site was recruited, resulting in the addition of 16 new amino acids (LGRAGLAPLNVCGLE, single-letter notation), none of them capable of being phosphorylated.

accessory molecule in the activation of T cells by cellular antigens. The role of signal transduction was highlighted by the inability of the Δ CD43⁺ clone to respond to Daudi stimulation although it responded well to stimulation by anti-CD3 monoclonal antibody.

To confirm that CD43 can serve as an accessory molecule, we examined a series of 48 CD43⁻, 48 CD43⁺, and 48 Δ CD43⁺ By155.16-derived uncloned TCR-CD3⁺ cell lines. As shown in Fig. 3b, 19 of the CD43⁺ lines (40%), but none of the CD43⁻ and Δ CD43⁺ lines responded to Daudi stimulation. Statistical analysis of the three groups revealed that the mean IL-2 production induced by Daudi stimulation of the 48 CD43⁻ lines did not differ significantly from that of the 48 Δ CD43⁺ lines. The mean of the 48 CD43⁺ lines was, however, significantly greater than that of the other groups ($P < 0.001$, by Student's *t*-test). This supports the conclusion that in the presence of TCR-CD3, CD43 can increase the efficacy of T-cell activation by antigen-presenting cells.

Binding assays were used to examine the specificity of the CD43-mediated interaction between T cell and antigen-presenting cell. Immunoaffinity-purified CD43 molecules were adsorbed to 96-well plastic tissue culture plates and the ability of Daudi cells and control A20 murine B cells to adhere to the CD43-coated plates was assessed. A significantly higher percentage of Daudi cells than A20 cells adhered to the CD43-coated plates (Fig. 4a). Preincubation of Daudi cells with liposomes containing purified CD43 molecules abrogated this adhesion,

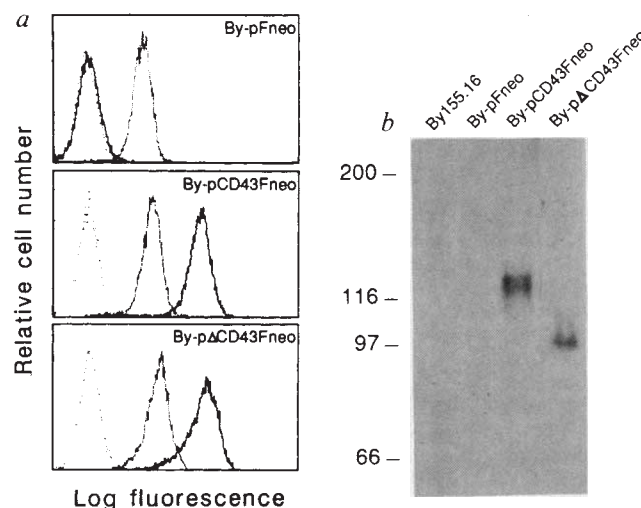


FIG. 2 Phenotype of transfected cells. *a*, Surface expression of the murine TCR (dotted line) and the extracellular domain of CD43 (solid line) on clones transfected with pFneo, pCD43Fneo and p Δ CD43Fneo was assessed by indirect immunofluorescent staining using F23.1 (anti-TCR monoclonal antibody) or L10 (anti-CD43 monoclonal antibody) and FITC-conjugated goat anti-mouse immunoglobulin antibody. The anti-immunoglobulin antibody was used alone as a control (highly dotted line). The three TCR-CD3⁺ clones shown were obtained by limiting dilution cloning. Not shown are the CD43⁻, CD43⁺, and Δ CD43⁺ clones which were TCR-CD3⁻. *b*, The relative molecular mass (M_r) of the proteins was determined by western blot analysis using L10 monoclonal antibody.

METHODS. The plasmids described in Fig. 1 were linearized with *EcoRI*, purified, and quantitated. By155.16 cells (2×10^7 ml⁻¹) were transfected by electroporation with one of the three plasmids (20 μ g ml⁻¹) using standard protocols²⁵. To generate clones, transfected cells were cultured at an initial density of 10^5 cells ml⁻¹ of RPMI 1640 containing 20% heat inactivated FBS, 100 U ml⁻¹ penicillin, 100 μ g ml⁻¹ streptomycin, and 50 μ M 2-mercaptoethanol (20% FBS-complete medium). Forty-eight hours later, cells were transferred to 10% FBS-complete medium containing 3 mg ml⁻¹ of G418. After 7 days of drug selection, cells were cloned by limiting dilution in 10% FBS-complete medium. Standard indirect immunofluorescent staining of clones to assess TCR and CD43 expression levels used F23.1 or L10 monoclonal antibody respectively. Western blot analysis of proteins used L10 monoclonal antibody.

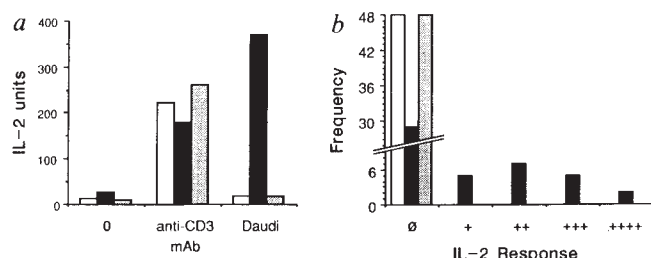


FIG. 3 Stimulation of transfected TCR-CD3⁺ cells. *a*, The CD43⁻ (□), CD43⁺ (■), and ΔCD43⁺ (▨) clones were stimulated with medium alone (○), the anti-CD3 monoclonal antibody 145-2C11, or Daudi cells. IL-2 production was measured as described below. Results from a single representative experiment repeated five times are shown. *b*, Individually derived CD43⁻ (□, *n*=48), CD43⁺ (■, *n*=48), and ΔCD43⁺ (▨, *n*=48) cell lines were assessed for their IL-2 production in response to Daudi cell stimulation. The IL-2 values were measured and scored on an increasing [0] to [++++] scale as defined below. The number of cell lines generating a given IL-2 response is shown on the ordinate.

whereas preincubation with liposomes containing purified HLA-DR had no effect (Fig. 4*a*). To demonstrate further the specificity of the CD43-Daudi cell interaction, wells coated with either CD43 or LFA-3 were preincubated with either anti-CD43 (B1B6 or G10-2, gifts of B. Axelsson and J. Ledbetter, respectively), anti-CD2 (TS2/18), or anti-LFA-3 (TS2/9) monoclonal antibodies. The plates were washed to remove unbound antibody, and Daudi cell adhesion assessed. B1B6 and G10-2 inhibited Daudi cell adhesion by 63% and 57%, respectively (Fig. 4*b*). In contrast, neither B1B6 nor G10-2 could inhibit the binding of a CD2⁺ hybridoma to plate-bound LFA-3, binding that could be inhibited by an anti-LFA-3 monoclonal antibody (TS2/9) (data not shown). Daudi binding to LFA-3 was not modified

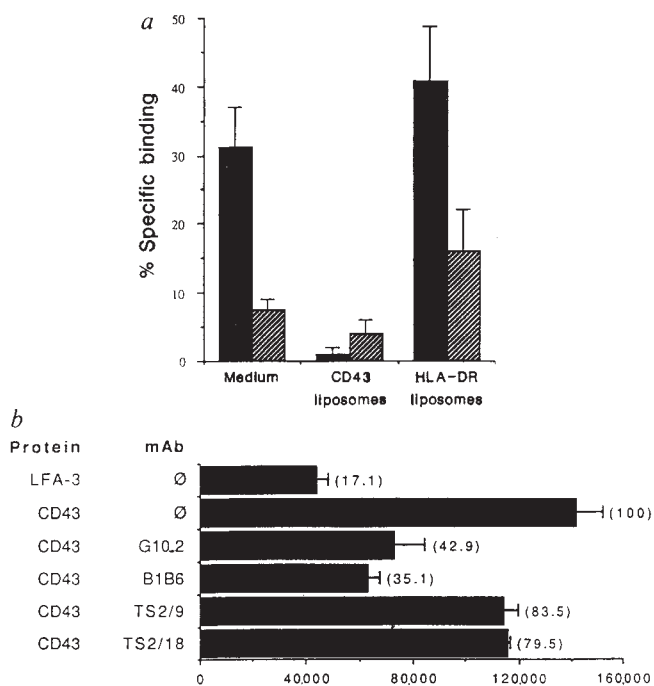
FIG. 4 Binding of human and murine B cells to human CD43. *a*, The specific binding of Daudi (■), a human B cell line, and A20 (▨), a murine B cell line, to CD43-adsorbed surfaces was assessed. Both cell types were preincubated with 10% FBS-complete medium containing either no liposomes, CD43-liposomes, or HLA-DR-liposomes. *b*, The binding of Daudi cells to CD43-adsorbed surfaces was inhibited by the anti-CD43 monoclonal antibodies B1B6 and G10-2. Data are expressed in c.p.m. The specific binding of Daudi cells to CD43 was assigned the value of 100%, and the binding of Daudi cells in the presence of monoclonal antibodies is expressed as a percentage of this binding (values in parentheses). The experiment shown is representative of five such experiments. Statistical analysis of all five assays revealed that the inhibition of Daudi cell binding to CD43 by anti-CD43 mAbs (B1B6 and G10-2) was significant ($P < 0.001$) and that the control anti-CD2 (TS2/18) and anti-LFA3 (TS2/9) monoclonal antibodies had no effect on binding ($P > 0.5$ and $P < 0.2$ respectively).

METHODS. CD43 was immunoaffinity purified from ¹²⁵I-labelled and unlabelled JY cell lysates using an L10 mAb-conjugated sepharose 4B column. PAGE of the isolated proteins followed by Coomassie blue staining and autoradiography, silver staining, or western blotting all revealed a single band of *M*_r 115,000, confirming the identity and purity of the sample. To a 96-well plate were added 50 μl of a 2.5 ng ml⁻¹ solution of purified CD43 in 0.3% octyl-β-glucoside, 0.1 M sodium bicarbonate buffer (pH 9.2). Equivalent amounts of immunopurified LFA-3²⁶ were added to control wells under identical conditions. Plates were incubated for 3 h at 37 °C, after which the remaining protein binding sites were blocked with 1% BSA for 2 h at 37 °C. Wells were extensively washed to remove unadsorbed proteins and [³H]-thymidine-labelled cells were then added to the plates at a density of 2×10^5 cells per well. The plates were centrifuged at 50 *g* for 1 min and placed in a 37 °C, 5% CO₂-in-air environment for 30 min. Unbound cells were then removed by aspirating the fluid contents of each well through an 18-gauge needle and carefully rinsing each well four times with 200 μl of prewarmed (37 °C) 10% FBS-complete medium. After washing, the radioactivity contained in each well was measured as c.p.m. Also determined were the total c.p.m. of cells initially added to the wells. The mean percentage of cells bound to BSA (alone)-adsorbed wells was 9.8 ± 4.7 (*n*=5 assays). Percentage specific binding was calculated as the: ((c.p.m. of cells retained by CD43-adsorbed wells) - (c.p.m. of cells retained by BSA-adsorbed wells)) ÷ ((total c.p.m. of cells added to the wells) - (c.p.m. of cells retained

METHODS. For the generation of nonclonal cell lines, cells were transfected as described in Fig. 2 and immediately transferred into 24-well tissue culture plates at a density of 10^5 cells per 1.5 ml of 20% FBS-complete medium. Forty-eight hours later, 0.5 ml of 12 mg ml⁻¹ G418 was added to each well to achieve a final concentration of 3 mg ml⁻¹. After 7–10 days of drug selection, viable cell lines were chosen and cultured in G418-free 10% FBS-complete medium for an additional 7–10 days. Cells from each cell line were then analysed for surface expression of CD3 and CD43 as described for the clones. Stimulation of clones and cell lines was performed in 24-well tissue culture plates containing 1.0 ml total reaction volume. Responder cells were used at a density of 5×10^4 cells ml⁻¹, (anti-CD3 monoclonal antibody was used at 1 μg ml⁻¹) and irradiated (50 Gy) Daudi cells were used at a density of 5×10^5 cells ml⁻¹. After 24 h stimulation, supernatants from cultured cells were assayed for IL-2 content by their ability to promote the proliferation of an IL-2 dependent murine T cell line, CTLL 20. Proliferation was assessed by measuring the incorporation of [³H]thymidine during the final 4 h of a 24-h incubation. Relative IL-2 units were determined for each supernatant by comparison with the proliferation of CTLL 20 supported by a standard rat concanavalin A supernatant, with half-maximal incorporation defined as 100 units. The cell lines were additionally assigned a response value based on the following scale: 0, <50 units; +, 50–80 units; ++, 80–110 units; +++, 110–150 units; +++, >150 units.

by the addition of anti-CD43 monoclonal antibodies (data not shown). These data strongly suggest that the CD43 molecule can bind to a ligand on the surface of Daudi cells. The avidity and specificity of this interaction are great enough to mediate the selective adherence of Daudi cells to immunoaffinity-purified CD43. In addition, this interaction can be inhibited by the anti-CD43 monoclonal antibodies B1B6 and G10-2, but not by control antibodies.

The data presented here demonstrate directly that in the presence of TCR-CD3 expression, CD43 expression by T cells enhances their response to antigen. The existence on antigen-presenting (Daudi) cells of a specific ligand for CD43 is strongly implied by the data. Following the engagement of a prospective



by BSA-adsorbed wells)). For the liposome inhibition assays, cells were preincubated for 1 h at 37 °C with 10% FBS-complete medium containing either no liposomes, CD43-liposomes, or HLA-DR-liposomes (1 μg protein per 20 nmol lipid) before plating. For the monoclonal antibody inhibition assays, protein-coated wells were incubated with a 1 μg ml⁻¹ monoclonal antibody solution for 4–6 h at 37 °C; excess monoclonal antibody was removed before the addition of cells. Before addition to protein-coated wells, Daudi cells were incubated with anti-CD18 monoclonal antibody (TS1/18) to inhibit homotypic adhesions.

ligand through its extracellular domain, CD43 probably transduces a costimulatory signal through its intracellular domain. This signal, in synergy with a subthreshold signal provided by the TCR-CD3, leads to T-cell activation. Although a ligand has not yet been isolated, the data presented here should lead to a greater understanding of the role of CD43 in the pathogenesis of the Wiskott-Aldrich syndrome and in the activation of T cells. □

Received 20 December 1990; accepted 4 March 1991.

- Gahmberg, C. G. & Andersson, L. C. *J. Biol. Chem.* **252**, 5888-5894 (1977).
- Gahmberg, C. G., Hayry, P. & Andersson, L. C. *J. Cell Biol.* **68**, 642-653 (1976).
- Brown, W. R. A., Barclay, A. N., Sunderland, C. A. & Williams, A. F. *Nature* **289**, 456-460 (1981).
- Remold-O'Donnell, E., Davis, A. E., Kenney, D., Bhaskar, K. R. & Rosen, F. S. *J. Biol. Chem.* **261**, 7526-7530 (1986).
- Carlsson, S. R. & Fukuda, M. *J. Biol. Chem.* **261**, 12779-12786 (1986).
- Axelsson, B., Hamnerstrom, S., Finne, J. & Perlmann, P. *Eur. J. Immun.* **15**, 427-433 (1985).
- Borche, L., Lozano, F., Vilella, R. & Vives, J. *Eur. J. Immun.* **17**, 1523-1526 (1987).
- Remold-O'Donnell, E., Zimmerman, C., Kenney, D. & Rosen, F. S. *Blood* **70**, 104-109 (1987).
- Parkman, R., Remold-O'Donnell, E., Kenney, D. M., Perrine, S. & Rosen, F. S. *Lancet* **ii**, 1387-1389 (1981).
- Axelsson, B., Youseff-Etemad, R., Hamnerstrom, S. & Perlmann, P. *J. Immun.* **141**, 2912-2917 (1988).
- Vargas-Cortes, M., Axelsson, B., Larsson, A., Berzins, T. & Perlmann, P. *Scand. J. Immun.* **27**, 661-671 (1988).
- Beyers, A. D., Barclay, A. N., Law, D. A., He, Q. & Williams, A. F. *Immun. Rev.* **111**, 59-77 (1989).
- Nong, Y.-H., Remold-O'Donnell, E., LeBien, T. W. & Remold, H. G. *J. exp. Med.* **170**, 259-267 (1989).
- Wilken, M., Bjork, P., Axelsson, B. & Perlmann, P. *Scand. J. Immun.* **29**, 363-370 (1989).
- Mentzer, S. J. et al. *J. exp. Med.* **165**, 1383-1392 (1987).
- Pallant, A. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1328-1332 (1989).
- Shelley, C. S. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 2819-2823 (1989).
- Sleekman, B. P. et al. *Nature* **328**, 351-353 (1987).
- Chatila, T. A. & Geha, R. S. *J. Immun.* **140**, 4308-4314 (1988).
- Axelsson, B. & Perlmann, P. *Scand. J. Immun.* **30**, 539-547 (1989).
- Piller, V., Piller, F. & Fukuda, M. *J. Biol. Chem.* **264**, 18824-18831 (1989).
- Bierler, B. E., Sleekman, B. P., Ratnofsky, S. E. & Burakoff, S. J. *A. Rev. Immun.* **7**, 579-599 (1989).
- Springer, T. A. *Nature* **346**, 425-434 (1990).
- Ohashi, P. S. et al. *Nature* **316**, 606-609 (1985).
- Potter, H. in *Current Protocols in Molecular Biology* (eds Ausubel, F. M. et al.) 9.3.1-9.3.2 (Wiley, New York, 1987).
- Takai, Y., Reed, M. L., Burakoff, S. J. & Herrmann, S. H. *Proc. natn. Acad. Sci. U.S.A.* **84**, 6864-6868 (1987).

ACKNOWLEDGEMENTS. This work was supported by the NCI (S.J.B.), IAD (F.S.R.) and GMI (E.R.-O.). J.K.P. is the recipient of a Physician-Scientist Award and a Medical-Scientist Award. B.E.B. is the recipient of a Clinician-Scientist Award from the American Heart Association and a McDonnell Scholar Award from the J. S. McDonnell Foundation. We thank W. C. Hahn for support and helpful discussions.

HIV enhancer activity perpetuated by NF- κ B induction on infection of monocytes

F. Bachelierie, J. Alcamí, F. Arenzana-Seisdedos & J.-L. Virelizier*

Unité d'Immunologie Virale, Institut Pasteur, 28 rue due Dr Roux, 75724 Paris Cedex 15, France

PERMISSIVENESS to replication of human immunodeficiency virus (HIV) differs in T lymphocytes and macrophages. In T cells, HIV transcription is poorly detected *in vivo*¹. Cloned, normal T lymphocytes show very little, if any, basal activity of the HIV enhancer and low nuclear expression of NF- κ B², a potent transcriptional activator of the HIV enhancer³⁻⁵. In contrast, fixed tissue macrophages express detectable HIV proteins, indicating permanent virus transcription⁶. One explanation for the perpetuation of virus infection in macrophages could be sustained nuclear NF- κ B expression. However, the U937 monocytic cell line, which is fully permissive to HIV replication, is known to express only low levels of nuclear NF- κ B⁷. We show here that chronic HIV infection results in both induction of a nuclear factor with antigenic properties indistinguishable from those of NF- κ B and permanently increased HIV enhancer activity. This phenomenon, which is independent of tumour necrosis factor, is associated with HIV

replication, and is thus likely to explain at least in part the perpetuation of HIV infection in monocytes.

To test the hypothesis that HIV replication induces a functional transcription factor(s) which can perpetuate HIV genome transcription, we analysed the activity of HIV enhancer-driven vectors in chronically infected U937 cells. Because the basal expression of nuclear NF- κ B varies according to experimental conditions⁸, we maintained cell cultures in the exponential growth phase and avoided all endotoxin or mycoplasma contamination. Figure 1a shows reverse transcriptase (RT) production by U937 cells which began at day 11 in one representative experiment (of 10) shown. Neutralization of input virus by soluble CD4 (sCD4) produced a 20-day delay in virus replication. Figure 1b shows that nuclear protein(s) were detected in infected cells which bound to an oligonucleotide with the sequence of the HIV enhancer, but not to an oligonucleotide with a mutation which abolishes NF- κ B binding³. This activity was barely detectable in uninfected cells, although a longer exposure of autoradiographs revealed a low-intensity pattern

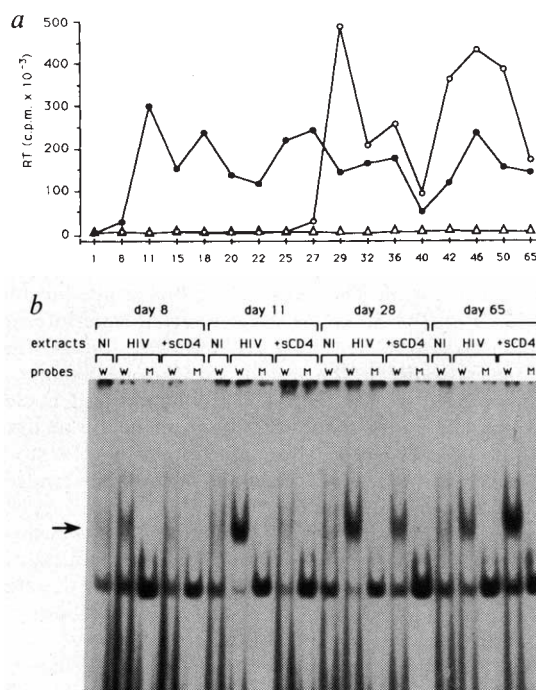


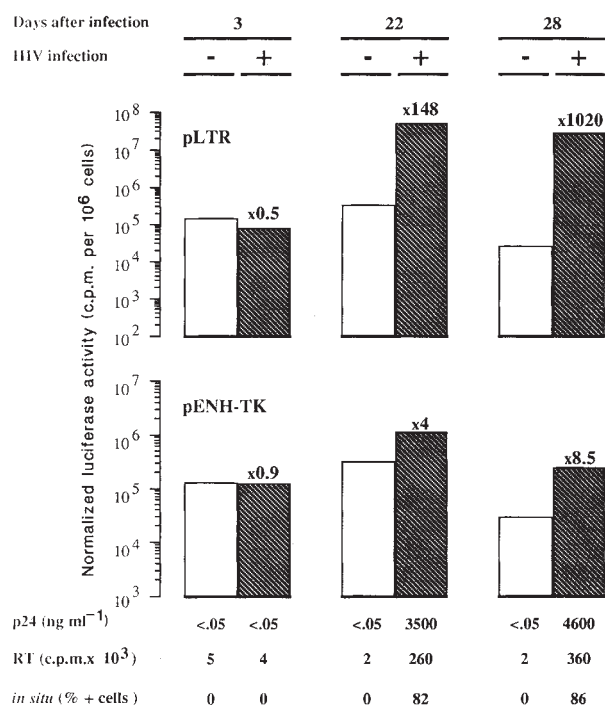
FIG. 1 Induction of nuclear HIV enhancer-binding proteins during HIV infection of monocytic U937 cells. **a**, Kinetics of reverse transcriptase (RT) activity in supernatants of U937 cells infected with HIV in the absence (●) or presence of soluble recombinant CD4 (American Biotechnology) (sCD4) (○) or mock-infected (Δ). The U937 cell line showed the following phenotype: Fcγ-receptor-, CD11a-, CD11b-, CD4-positive, CD3-, CD14-, CD25-negative and non-adherent. Cells were incubated for 1 h with a preparation of HIV LAV-1 Bru strain grown in U937 which was clarified, filtered (0.2 μm pore size) and used at a concentration of 6×10^5 c.p.m. of RT per 1×10^6 cells per ml. Mock-infected cells received RPMI-1640 medium supplemented with 5% FCS. No polybrene was added. Neutralization with sCD4 ($50 \mu\text{g ml}^{-1}$) of 1 ml virus stock (6×10^6 c.p.m. of RT) was performed for 1 h at 4°C before addition to 9 ml cell suspension for 1 h at 37°C. All cells were then washed three times with PBS and maintained at 37°C, in a 5% CO₂ atmosphere at a concentration of 2.5×10^5 per ml of culture medium (RPMI supplemented with 10% FCS). RT activity (c.p.m. per 50 μl) was measured by standard microassay¹⁷. **b**, Bandshift assays with nuclear extracts (5 μg) from infected cells in the absence (HIV) or presence of sCD4 (+sCD4) or control cells (NI) at days 8, 11, 28 and 65 after infection were prepared as previously described^{2,18}. The binding probes were end-labelled double-strand synthetic oligonucleotides corresponding to the wild-type NF- κ B sequence of the HIV enhancer (W) (gift from C. Auffray and A. Delaroché): 5'-ACAAGG-GACTTTCGCTGGGACTTTCAGGGA-3', or the mutated HIV enhancer sequence³ (M): 5'-ACAACCTCACTTTCGCTGCTCACTTTCAGGGA-3' where mutated sites are underlined. Arrow denotes specific binding. Autoradiography was for 5 h.

* To whom correspondence should be addressed.

FIG. 2 Activity of the HIV LTR- or HIV enhancer-TK-luciferase expression vectors transfected into infected (■) or control (□) U937 cells at days 3, 22 and 28 after infection. The HIV LTR-luciferase vector was the *Xho*I (−640)–*Hind*III (+78) fragment containing the U3 + R regions from the LAV-1 Bru strain cloned into the pC-luc plasmid¹⁹ generating pLTR, as described elsewhere¹⁹. The Enh-TK-luciferase vector containing the HIV-enhancer sequence was constructed by cloning the 270-basepair (bp) *Bgl*II (Klenow)–*Sac*I fragment of plasmid pHL-03TK-CAT¹⁸ into the *Hind*III (Klenow)/*Sac*I sites of the pC-luc polylinker¹⁹. This vector contains three copies of the synthetic HIV-enhancer oligonucleotide described in Fig. 1 as (W), cloned upstream of the weak, truncated Herpes simplex thymidine kinase (TK) promoter (−105–+51). Transfection was performed according to a modified DEAE-dextran technique²⁰ using 5 µg plasmid DNA per 1.5×10^7 cells. Luciferase activity was measured¹⁹ in cell lysates made 60 h after transfection and is expressed as c.p.m. per 1×10^6 living cells, after normalization in function of the relative activity of a pTK-luciferase vector transfected into control (□) or infected (■) cells. Comparable amounts of transfected plasmid were found by DNA Hirt extraction²¹ 2 h after transfection in control and infected cells and input DNA plasmid correlated with luciferase expression of the TK-luciferase plasmid. The pTK-luciferase vector was constructed by ligation of the *Pvu*II–*Bgl*II fragment containing the TK promoter from pBL-CAT²² vector with the *Pvu*II–*Bam*HI fragment of pC-luc. Numbers at the top of bars represent fold amplification of luciferase activity in infected cells compared to that expressed by the same vector in control cells in the same experiment. RT assay was performed and expressed as described in Fig. 1. Levels of p24 in culture supernatants are indicated as nanograms per millilitre, based on p24 standards (NEN Research product). For *in situ* detection^{23,24}, an HIV (LAV-1 Bru) *gag* insert (*Sac*I (224)–*Hind*III (2078)) was cloned into a pSP72 vector which provides SP6/T7 promoters (Promega-Biotech). RNA probes were generated with either Sp6 (sense) or T7 (antisense) RNA polymerase and hybridized *in situ* with control or infected cells. Results are expressed as % of positive cells with the T7 (antisense) *gag* probe. When

similar to that in Fig. 4. This NF- κ B-like factor was not seen in the nuclei of U937 cells transiently protected from infection by sCD4, but was observed (day 28) following escape from protective neutralization of the input virus. Chronically infected cells remained positive for both RT production and nuclear NF- κ B expression for the ten months duration of this analysis. These results were observed after infection with HIV stocks grown in either CEM or U937 cells and with either crude or sucrose gradient-purified virus.

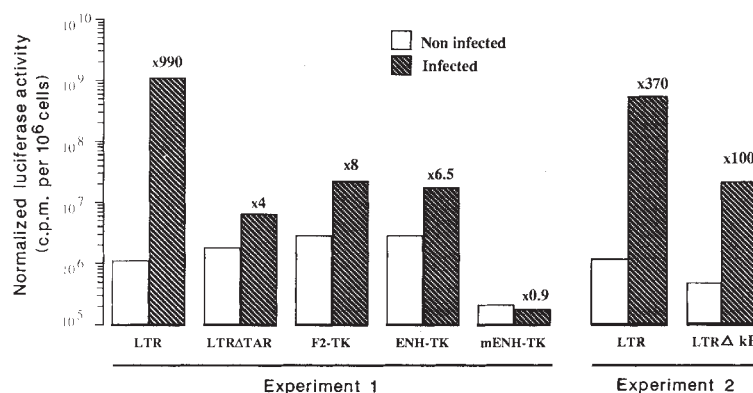
We recently observed in human T cell clones, that translocation of NF- κ B into the nucleus is not necessarily sufficient to induce transcription dependent on HIV enhancer². We therefore transiently transfected chronically infected cells with a series of luciferase expression vectors under the control of various regions of the LAV-1 Bru long terminal repeat (LTR). HIV-infected cells showed enhanced activity of an expression vector under the control of three copies of an oligonucleotide with the sequence of the two 10-base pair direct repeats of the HIV enhancer (pENH-TK) (Fig. 2). This was only seen after detectable virus replication (day 22) at a stage when more than 80% of cells were actively producing HIV *gag* transcripts as shown



positive cells were observed, the percentage was evaluated on at least 500 cells. Negative results were recorded after screening of at least 10^4 cells.

in the two representative *in situ* hybridization experiments performed 22 and 28 days after the onset of infection. Strong NF- κ B like nuclear binding activity was detected in infected, but not in control, cells on day 22 (Fig. 4a). Transactivation in cells replicating HIV was similar to that in control cells stimulated with phorbol myristate acetate (PMA) (4–15-fold and 1.3–9.6-fold amplifications, respectively). The activity of a vector under the control of the whole (U3 + R) HIV LTR (pLTR) in infected cells was much greater than that of the enhancer-driven vector, probably reflecting HIV *tat* protein function, because deletion of the TAR sequence led to a response comparable in intensity to that of the enhancer-TK-luciferase vector (Fig. 3). Similarly, an LTR fragment (−158–69) including the HIV enhancer motif and flanking sequences cloned upstream of the TK promoter (pF2-TK) transfected into infected cells showed an increased activity, comparable with that of the enhancer-driven vector. Such increased activity was not seen in infected cells transfected with a vector under the control of an oligonucleotide mutated at bases critical for binding of NF- κ B (Fig. 3). Therefore, the protein factor in chronically infected cells recognizes the NF- κ B consensus motif and is functional. Interestingly, a modest but

FIG. 3 Activity of LAV-1 Bru LTR luciferase expression vectors transfected into infected (■) or control (□) U937 cells. The ENH-TK luciferase construct is described in Fig. 2. The HIV LTR vector contains the *Bgl*II (−485)–*Hind*III (+78) fragment. The LTRΔTAR vector contains the *Bgl*II (−485)–*Sac*I (+38) LTR fragment deleted of the TAR region and cloned into the pC-luc plasmid. The F2-TK-luciferase vector was the (−158–69) *Hae*III fragment of the HIV LTR¹⁸ containing the HIV-enhancer sequence cloned upstream of the TK promoter. The mENH-TK luciferase construct contains four copies of a mutated synthetic HIV-enhancer oligonucleotide (described in Fig. 1 as M) cloned upstream the TK promoter. The LTRΔκB was obtained by insertion into the pC-luciferase plasmid¹⁹ of the κ B consensus motifs-deleted *Xho*–*Hind*III fragment of the Lav-1 Bru LTR from the original CAT construct of A. Rabson²⁵. Transfection, luciferase activity detection and normalization were performed as described in Fig. 2. Numbers at the top of bars represent fold amplification of luciferase activity in infected cells compared to that



expressed by the same vector in control cells in the same experiment.

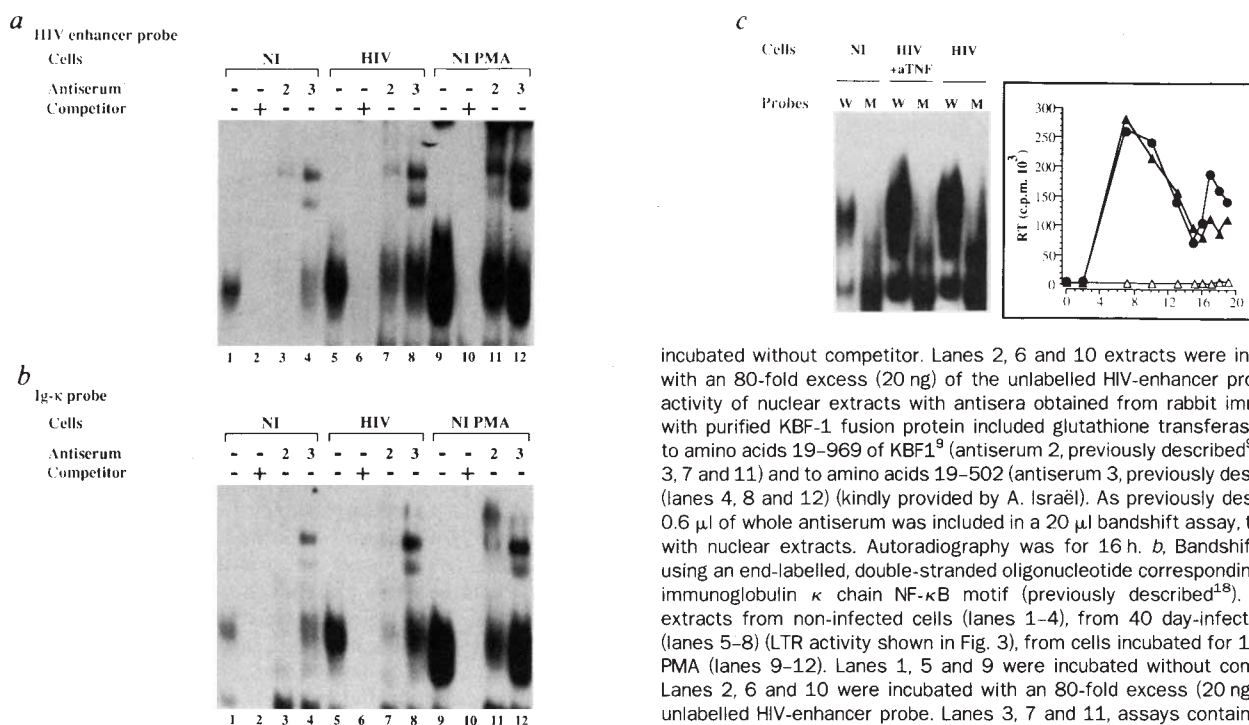


FIG. 4 Binding properties of specific nuclear factor(s) detected in uninfected and infected cell extracts. Assays conditions as described in Fig. 1. **a**, Bandshift assay was performed using the Klenow-labelled, double-stranded oligonucleotide corresponding to the HIV-enhancer sequence. Nuclear extracts (2.5 μ g) were made from uninfected cells (lanes 1–4), from 22 day-infected cells (lanes 5–8) (LTR activity shown in Fig. 2) and from cells incubated for 16 h with PMA (lanes 9–12). Lanes 1, 5 and 9 extracts were

incubated without competitor. Lanes 2, 6 and 10 extracts were incubated with an 80-fold excess (20 ng) of the unlabelled HIV-enhancer probe. Reactivity of nuclear extracts with antisera obtained from rabbit immunized with purified KBF-1 fusion protein included glutathione transferase linked to amino acids 19–969 of KBF1⁹ (antiserum 2, previously described⁹) (lanes 3, 7 and 11) and to amino acids 19–502 (antiserum 3, previously described⁹) (lanes 4, 8 and 12) (kindly provided by A. Israël). As previously described⁹, 0.6 μ l of whole antiserum was included in a 20 μ l bandshift assay, together with nuclear extracts. Autoradiography was for 16 h. **b**, Bandshift assay using an end-labelled, double-stranded oligonucleotide corresponding to the immunoglobulin κ chain NF- κ B motif (previously described¹⁸). Nuclear extracts from non-infected cells (lanes 1–4), from 40 day-infected cells (lanes 5–8) (LTR activity shown in Fig. 3), from cells incubated for 16 h with PMA (lanes 9–12). Lanes 1, 5 and 9 were incubated without competitor. Lanes 2, 6 and 10 were incubated with an 80-fold excess (20 ng) of the unlabelled HIV-enhancer probe. Lanes 3, 7 and 11, assays contained antiserum 2 and lanes 4, 8 and 12 received antiserum 3. **c**, (left panel) Bandshift assay of a Klenow-labelled double-stranded oligonucleotide corresponding to the HIV-enhancer sequence (W) or a mutated HIV enhancer sequence (M) using nuclear extracts made on day 19 after infection from infected cells cultured with 1 μ g per ml of neutralizing anti-TNF monoclonal antibody (HIV + α TNF) or with a control antiserum (HIV) or from uninfected cells (NI). Right panel, kinetics of reverse transcriptase (RT) activity in supernatants of U937 cells infected with HIV (LAV-1 Bru strain) cultivated in presence of antibody to TNF ($\blacktriangle$), in presence of a control antiserum ($\bullet$) or mock-infected ($\triangle$).

consistent difference in activity was seen between the enhancer- and the mutated oligonucleotide-driven vectors in control, uninfected cells. Therefore, the binding activity in control cells after sufficient autoradiographic exposure (see Fig. 4a and b) is also functional.

The bandshift assays in Fig. 4 indicate that the HIV enhancer binding complex was specifically competed by an oligonucleotide with the NF- κ B consensus motif in the context of the HIV enhancer itself and detected with both an HIV enhancer (Fig. 4a) and an immunoglobulin κ chain enhancer probe (Fig. 4b). Addition of specific antibody to purified recombinant p50⁹ caused a profound decrease and partial upshift (depending on the antiserum used) of the binding observed in nuclear extracts from infected, uninfected and PMA-stimulated cells. We have confirmed (A. Israël, personal communication) that antiserum number 2 does not immunoprecipitate the c-rel protein translated *in vitro* from c-rel DNA (gift from N. Rice). Preliminary experiments have also shown that the binding observed in infected cell nuclei is not influenced by an antiserum to the C-terminal part of the c-rel protein (gift from I. Verma and L. Kerr) which does not show homology with the p50 protein of the NF- κ B complex. Thus, the protein complex showed migration and antigenic properties indistinguishable from those of NF- κ B^{9,10}.

The mechanisms through which HIV infection leads to NF- κ B induction in U937 cells is not clear. Contamination of our viral preparations by non-viral products such as lymphokines or endotoxins was controlled by sCD4 neutralization assays. Direct induction by binding of viral particles to membrane CD4, a transducing molecule^{11–13}, is unlikely because excess sucrose gradient-purified, viable LAV-1 Bru particles did not induce NF- κ B 5–48 h (not shown) or 8 days (see Fig. 1) after infection. Autocrine secretion of tumour necrosis factor (TNF) was also

ruled out (Fig. 4c). TNF-neutralizing antibody in the culture medium did not influence RT production or NF- κ B induction. Moreover, TNF and interleukin 1 activities were not detected (not shown). The mechanism(s) of NF- κ B induction during chronic infection of U937 cells is unknown, but seems to be associated with HIV replication. For example, proteolysis may regulate NF- κ B binding activity¹⁴ during HIV infection because others (A. Israël, personal communication) have observed that the HIV protease influences the maturation and potentially the translocation, of NF- κ B.

We have shown that HIV infection itself upregulates a low constitutive nuclear expression of NF- κ B in U937 cells. Therefore, HIV infection may be self-perpetuating in monocytic cells. HIV latency is presumably the result of low expression of nuclear factors critically involved in LTR-dependent transcription¹⁵. Following *tat* production after HIV LTR-activation, an amplification loop allows full HIV transcription and replication. Recent evidence indicates that the HIV enhancer domain is critical for HIV *tat* functioning¹⁶. Figure 3 (experiment 2) provides further evidence that *tat* function is decreased when NF- κ B consensus motifs are deleted in the LTR. The activity of the $\Delta\kappa$ B vector compared to that of the wild-type LTR was decreased by a factor of 20 (19×10^6 versus 390×10^6 c.p.m. per 10^6 cells of normalized luciferase activity, respectively) in infected cells. This is consistent with the finding²⁶ that replication in U937 of a $\Delta\kappa$ B-deleted provirus is diminished but not abolished. These data indicate that the κ B sequences of the LTR participate clearly, although not solely, in the optimal transactivating effects of *tat*. Therefore HIV replication itself, through induction of an increased HIV enhancer activity, may initiate and perpetuate the *tat*-mediated amplification of HIV genome transcription, thus modifying the monocyte cell environment to its own advantage. □

Received 5 December 1990, accepted 6 March 1991.

- Poli, G. et al. *Science* **244**, 575–577 (1989).
- Hazan, U. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 7861–7865 (1990).
- Nabel, G. & Baltimore, D. *Nature* **326**, 711–713 (1987).
- Siekewitz, M. et al. *Science* **238**, 1575–1578 (1987).
- Kaufman, J. D. et al. *Molec. cell. Biol.* **7**, 3759–3766 (1987).
- Gendelman, H. E. et al. *AIDS* **3**, 475–495 (1989).
- Griffin, G. E. et al. *Nature* **339**, 70–73 (1989).
- Griffin, G. E. et al. *Nature* **343**, 219 (1990).
- Kieran, M. et al. *Cell* **62**, 1007–1018 (1990).
- Ghosh, R., et al. *Cell* **62**, 1019–1029 (1990).
- Merrill, J. E., Koyanagi, I. & Chen, I. S. *J. Virol.* **63**, 4404–4408 (1989).
- Poli, G. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 782–785 (1990).
- Wahl, L. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 621–625 (1989).
- Gilmore, T. D. *Cell* **62**, 841–843 (1990).
- Pomerantz, R. J., Trono, D., Feinberg, M. B. & Baltimore, D. *Cell* **61**, 1271–1276 (1990).
- Berkhout, B., Gatignol, A., Rabson, A. B. & Jeang, K.-T. *Cell* **62**, 757–767 (1990).

- Schwartz, O., Henin, Y., Marechal, V. & Montagnier, L. *AIDS Res. hum. Retrovir.* **4**, 441–448 (1988).
- Israël, N. et al. *J. Immun.* **143**, 3956–3960 (1989).
- Schwartz, O., Virelizier, J.-L., Montagnier, L. & Hazan, U. *Gene* **88**, 197–205 (1990).
- Wang, Y., Larsen, A. S. & Peterlin, B. M. *J. Exp. Med.* **166**, 625–637 (1987).
- Hirt, R. J. *Molec. Biol.* **26**, 365–369 (1967).
- Luckow, B. & Schütz, G. *Nucleic Acids Res.* **15**, 5490 (1987).
- Bachelier, F. et al. *J. Virol.* **69**, 3059–3062 (1990).
- Haase, A., Brahic, M., Stowing, L. & Blum, H. *Methods Virol.* **7**, 189–226 (1984).
- Du, E. J., Maury, W. J., Folks, T. M., Fauci, A. S. & Rabson, A. B. *Proc. natn. Acad. Sci. U.S.A.* **86**, 5974–5978 (1989).
- Harrich, D., Garcia, J., Mitsuyasu, R. & Gaynor, R. *EMBO J.* **9**, 4417–23 (1990).

ACKNOWLEDGEMENTS. We thank A. and N. Israël, F. Logeat and S. Michelson for helpful discussions. A. Israël for antibody number 2 and 3 to p50⁹ and C. Auffray for oligonucleotide synthesis. We thank D. Thomas and F. Aillet for technical assistance. The Agence Nationale de Recherche sur le Sida (A.N.R.S.) supported this work. We are grateful for a fellowship from A.N.R.S. to F.B. and from Caja de Ahorros de Madrid (Spain) to J.A.

Regulation of cellular responsiveness to inductive signals in the developing *C. elegans* nervous system

David A. Waring & Cynthia Kenyon

Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143-0554, USA

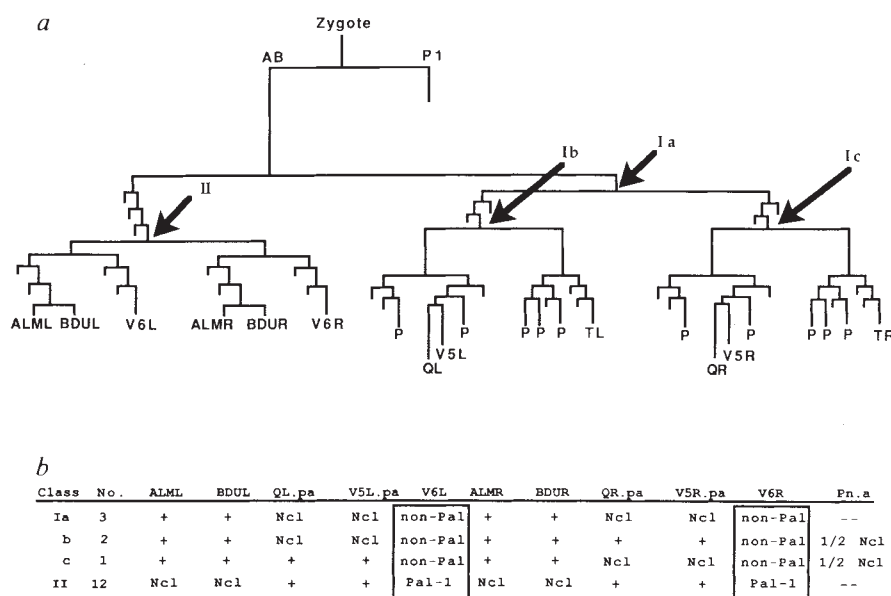
IN *Caenorhabditis elegans*, cell–cell communication is required to form a simple pattern of sensory ray neurons and cuticular structures (alae). The *C. elegans pal-1* gene initiates one developmental pathway (ray lineages) simply by blocking a cell–cell interaction that induces an alternative pathway¹. Here we show by mosaic analysis that *pal-1*⁺ acts by preventing specific cells from respond-

ing to inductive signals. The results indicate that although cell signals play a critical role in generating this pattern, they do not provide spatial information. Instead, signals are sent to many, if not all, of the precursor cells, and the ability to respond is spatially restricted. This patterning strategy thus differs from many well known models for pattern formation in which localized inductive signals influence a subset of cells within a field. We find that *pal-1* encodes a homeodomain protein and so is likely to regulate transcription. The *pal-1*⁺ protein could block the response to cell signals either by repressing genes involved in signal transduction or by acting directly on downstream genes in a way that neutralizes the effects of the intercellular signals. Genetic experiments indicate that one candidate for such a downstream gene is the *Antennapedia*-like homeotic selector gene *mab-5*.

In a newly hatched *C. elegans* male, seven cells (V1–V6 and T) lie in a row along each side of the animal². Signals from neighbouring V cells induce V cells located in anterior body regions to generate epidermal lineages ('alae lineages'), including cells that produce alae³. Posterior V and T cells generate

FIG. 1 Mosaic analysis of *pal-1*. *a*, Lineage diagram showing the relationships between V5, V6, T and the cells in which the Ncl-1 phenotype was used to score for the presence of the free duplication (see methods below)⁹. The arrows identify the point of loss of the duplication in various classes of mosaics. In the cases where the point of duplication loss is ambiguous, the arrow points to the last cell in which the loss could have occurred. The diagram represents only lineal relationships between cells, and not the relative timing of cell divisions. *b*, Summary of the Ncl and *Pal-1* phenotypes observed in different classes of mosaic animals.

METHODS. Mosaic animals arose spontaneously in the strain *sDp3; pal-1 (e2091) dpy-17(e164) ncl-1(e1865) unc-36(e251); him-5(e1490)*. The free duplication *sDp3*, which carries wild-type copies of these linked genes, is lost in roughly 1/400 cell divisions^{4,10}. The mutation *ncl-1(e1865)* causes cells to have unusually large nuclei. This gene has been shown to act cell-autonomously and thus allows one to determine cell genotypes (E. Hedgecock, personal communication). The Ncl phenotype is especially visible in neuronal cells. Because we were not able to score the Ncl-1 phenotype in all cells directly, the genotype of some cells was deduced by scoring linearly related cells in order to determine in which cell the duplication was lost during development. Classes Ia, Ib, and Ic: these animals were identified on the basis of the Ncl-1 phenotype of Q-derived neurons (Q.paa and Q.pap), the P(1–10).a-derived cells and the V5.pa descendants (the postdeirid). Animals were generally examined in early L3 for the Ncl-1 phenotype and then scored as young adults for the *Pal-1* phenotype. Animals were classified as class Ia mosaics if the Q-derived cells and postdeirid cells on both sides of the animal were Ncl and the ALM and BDU cells on both sides were non-Ncl. Animals were placed in class Ib or Ic if the ALM and BDU cells on both sides were non-Ncl, Q-derived cells and the postdeirid cells on one side were Ncl and on the other side were non-Ncl, and Pn.a-derived cells from one of each bilateral pair were Ncl (either P1 or P2, P3 or P4, and so on). Class II: two class II animals were identified by



scoring the Ncl phenotype in L3 animals. The remaining class II animals were found as adult males exhibiting a *Pal-1* phenotype (both sides) but not a *Dpy Unc* phenotype (indicating that they carried the duplication in some cells), and then scored for the Ncl phenotype. Eleven such animals were identified. Ten of these had Ncl ALM and BDU cells (both sides), indicating a loss between ABa and ABapp. One animal did not exhibit a Ncl phenotype (we scored neurons and coelomocytes), and so may have been a recombinant or a double loss. The absence of other Ncl patterns among these phenotypically *Pal-1* animals is significant, because it suggests that losses in other lineages do not produce a *Pal-1* phenotype. But it should be noted that we have not tested all lineages explicitly. One animal scored as *Pal* had a partial *Pal-1* phenotype: one side had no V6 rays while the other had 3 V-derived rays. Such partial phenotypes are found in ~15% of the animals in a *pal-1* homozygous strain¹.

different lineages ('ray lineages') that consist of ray neuroblasts as well as epidermal cells². The gene *pal-1* (for 'posterior alae') is required for V6 to generate its characteristic ray lineage instead of alae lineages. The V6 ray-to-alae transformation in *pal-1* mutants is the consequence of inappropriate cell communication: ablation of V6's neighbour, T, allows V6 to generate its normal ray lineage in spite of the *pal-1* mutation. Thus the function of wild-type *pal-1* is to block or override signals from T that would otherwise induce V6 to produce alae lineages (see Fig. 4a). This implies that pattern formation among these cells involves the localized inhibition of cell interactions¹.

Wild-type *pal-1* activity could prevent V6 from communicating effectively with T either by preventing T from sending signals or by preventing V6 from receiving signals. To distinguish between these possibilities, we generated genetic mosaic animals that were genotypically *pal-1*⁺ in some cells and *pal-1*⁻ in others. These animals were isolated from a strain that carried a *pal-1*⁻ mutation on each chromosome and a wild-type *pal-1*⁺ gene copy on a free duplication. During mitosis this free duplication (and thus the *pal-1*⁺ gene) is lost spontaneously at a low frequency, creating clones of *pal-1*⁻ cells in otherwise *pal-1*⁺ animals. Using the previously described markers *unc-36*, *dpy-17* and *ncl-1* (refs

4 and 5; E. Hedgecock, personal communication) as lineage markers, we could identify mosaic animals in which the *pal-1* genotypes of V6 and T were inferred to be different (Fig. 1). In class I mosaics T had a mutant *pal-1*⁻ genotype, whereas V6 had a *pal-1*⁺ genotype. In all these animals V6 developed normally (that is, it produced ray lineages). This result indicates that *pal-1*⁺ activity is not required in T. In class II mosaics V6 had a mutant *pal-1*⁻ genotype, whereas T had a *pal-1*⁺ genotype. In all of these animals, V6 had a mutant Pal-1 phenotype (produced alae lineages). This result indicates that *pal-1* acts in V6 or in close lineal relatives of V6 that were also mutant in class II mosaics. Because none of these relatives of V6 is located in the posterior body region, it seems unlikely that *pal-1* functions in these cells. Thus we believe that *pal-1*⁺ acts in the responding cell, V6, to prevent effective cell-cell communication.

This mosaic analysis indicates that T sends a signal to V6 in the wild type, even though V6 cannot respond to it. Thus, in this system, spatial information is not imparted by targeting alae-inducing signals to one type of precursor cell. Rather, alae and ray precursors are both exposed to the signals, and the pattern is established by regulating the cell's ability to respond.

FIG. 2 The nucleotide and deduced amino-acid sequence of a *pal-1* cDNA. The arrow indicates the AUG most likely to be the translation start site on the basis of sequences of other *C. elegans* start sites (M. Perry, G. Hertz and W. Wood, personal communication). The homeodomain is underlined. The homeodomain is most similar to the homeodomains of the *Drosophila* gene *caudal* and the mouse gene *cdx-1*. The *pal-1* homeodomain was named *ceh-3* by Burglin *et al.*, who first reported these homologies¹¹. The significance of these homologies is unclear. We have found a region of homology between *pal-1* and *mab-5* in addition to the homeodomain. In both genes there is encoded a serine-rich region followed by SAAAAAAN (in single-letter amino-acid code; shown in box). Such polyamino-acid stretches are common in homeodomain proteins, but their function is unknown¹². That both *pal-1* and *mab-5* act in V6 to promote ray formation makes this homology possibly significant.

METHODS. The *pal-1* gene was identified by transformation rescue as described¹³. The *pal-1* gene maps genetically close to the gene *ced-4* which has been cloned (J. Yuan and H. R. Horvitz, personal communication; H. Madhani, and C.K., unpublished results). Cosmid clones from the region (kindly provided by A. Coulson) were co-injected into *pal-1*(e2091); *unc-31*(e169) hermaphrodites with a cosmid containing the wild-type *unc-31* gene, and the non-Unc progeny were scored for the Pal phenotype. The cosmid W05E6 fully rescued the Pal-1 phenotype. Burglin *et al.* cloned and sequenced a homeobox, *ceh-3*, which is located on W05E6 (ref. 11). Thinking that *ceh-3* might be in the *pal-1* gene, we cloned successively smaller fragments containing *ceh-3* and eventually obtained a 5.5-kb *PstI-KpnI* clone (pWPK6) which fully rescues the Pal-1 phenotype. (These transformations were done using as cotransformation markers either *unc-31* (R. Hoskins, personal communication) or the plasmid pRF4, which carries the dominant allele of *rol-6*, *su1000* (J. Kramer and C. Mello, personal communication).) Two independent cDNAs were obtained by hybridization to one of our rescuing clones. The cDNAs were identical in sequence, except that one clone was 7 base pairs longer at the 5' end of the gene and had three additional bases before the poly(A) tail. The sequence shown is that of the

```

1: .CA AGT TTG AGA ATG TCG GTC GAT GTC AAG TCG GAT TTT TCG GAA AAC GAG TCC TCA TCA ACA CCT TCA CCA ACA
   : ... ser leu arg met ser val asp val lys ser asp phe ser glu asn glu ser ser ser thr pro ser pro thr
76: ACA GTT CCA GCA GAT GTC ACG TGG CCT CAC TAT CCA ATG ATG CCA TTC ATG CAA CCA CAT CCT CTG AGA GAA AAG
   : thr val pro ala asp val thr trp pro his tyr pro met met pro phe met gln pro his pro leu arg glu lys
151: ATG CTG CAA CCA ACA TTT GAT CCA CAA ATC TAT GGA AGG TGG TCA CAA ATG GGT GAT ACT GGT TTT TAC GGA CAT
   : met leu gln pro thr phe asp pro gln ile tyr gly arg trp ser gln met gly asp thr gly phe tyr gly his
226: CCG GAT CTC TAT CCA TTC GGC CTT CCA CAA CTC GCT GCC AAC GGA CAA ATT CCA GCA GTA GAA GCC GTT GAT GTG
   : pro asp leu tyr pro phe gly leu pro gln leu ala ala asn gly gln ile pro ala val glu ala val asp val
301: AAG CCA CCG CTT TCG AAT GGA AGT AGC AGT AGT GAT AGT GGA ATG TAT CCA TCA CCG AGT GAT ATG ATG ACA CCG
   : lys pro pro leu ser asn gly ser ser ser ser asp ser gly met tyr pro ser pro ser asp met met thr pro
376: TTC CCA TCA ACA TCA TCT GGA GCT GCA TCT TCT TCA GAG CTT TCA GCA GCT GCT GCT GCT GCA GCC AAC TAC CAG
   : phe pro ser thr ser ser gly ala ala ser ser ser glu leu ser ala ala ala ala ala ala ala asn tyr gln
451: ATG AGA GCT GCC ACG TGT TAT CAG CAA TCT GTT TGG CCG TTT ATG GAT TAT CAA CAG TTT CAG GGA TTC TCG TGG
   : met arg ala ala thr cys tyr gln gln ser val trp pro phe met asp tyr gln gln phe gln gly phe ser trp
526: AAA ATG CCT CTT GGA AAC AAT CAT GGA AAG GAT CGA AGA TCA AGC AGT GAT GGG AAA ACA TTA CCA ACG GGA CCT
   : lys met pro leu gly asn thr his gly lys asp arg arg ser ser ser asp gly lys thr leu pro thr gly pro
601: GGA ACG AAT AAT GTC CGT GTA CGA ACA GCT GAC AAA TAC CGT ATG GTT TAC TCG GAT TAT CAA CGC CTT GAA CTG
   : gly thr asn asn val arg val arg thr ala asp lys tyr arg met val tyr ser asp tyr gln arg leu glu leu
676: GAA AAA GAA TTC CAT ACA TCG CCA TTC ATA ACA TCA GAT CGC AAG TCC CAA TTG TCG ACG ATG TTA AGC TTG ACA
   : glu lys glu phe his thr ser pro phe ile thr ser asp arg lys ser gln leu ser thr met leu ser leu thr
751: GAA CGA CAA ATC AAG ATT TGG TTT CAA AAT AGG CGT GCA AAG GAT CGT CGT GAC AAA CAG AAG ATT CGG CTA TAA
   : glu arg gln ile lys ile trp phe gln asn arg arg ala lys asp arg arg asp lys gln lys ile arg leu ***
826: GTA CTC ATC TAC TTA CAA AGA AAA TGT ACA AAA ATT CTG GGA ATT TTC ATA TTT TTC TAG AAT ATT GTA CAC ATT
901: TAT TAT ATT TTT ACA TTT CTC TCG CTT GAA TAC GGT TAA ATT CAC ACT TTT ATA TAA TTT CTT CCT GTC CCA TCT
976: CCC ATT CTC TCA CAT CAG AAT ATT CTA ATC ATC ACT GCC CTT AAC ATA TAT ATA TAT ATA TTT ATA ACT TTC CGA
1051: TCA TTC CTT CTG CTT TGT AAA CTA ATC CAA TCA TTC TCT AAT GAA TTA TTC CCT AAT ATA GAG CTT CTT TAT TTA
1126: TTG TGA TTT CTT GTT TCT TAC CCT TCT CAA GGT TTA GGA AAA AGT TTA GGT CCC CAT TTG TTA TGT ATT TAC TTC
1201: CCA CTC TGA ACG GAT CCA TTT TGA ACA TGT TTT CTA TTT TTT GGA AAA GTT TAG AGA ATT CTT TGC TTT TCT CAC
1276: ATA TCT TGA TTG CAT ATT TTT TAT AAC AAA ACA ATG AAA CAA TTT TGA AGA AT...poly-A

```

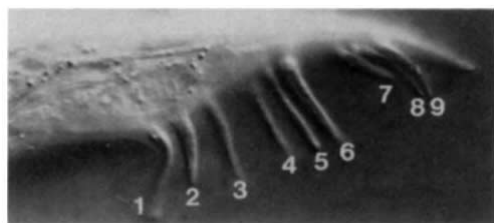
longer clone (sequenced on both strands using Sequenase 2.0). The first 10 bases of the longer cDNA (and the first three bases of the smaller cDNA) correspond to the last bases of one of the 19-base-pair *trans*-spliced leader sequences found at the 5' end of many *C. elegans* genes¹⁴. This leader sequence suggests that these cDNAs are, with the exception of the remaining bases of the *trans*-spliced leader, of full length. The genomic clone that rescues *pal-1*, (pWPK6), contains about 1 kb of sequence upstream of the 5' end of the cDNA and roughly 400 bases downstream of the 3' end of the cDNA. On several occasions an additional ray was present in animals transformed with pWPK6, presumably caused by overexpression or ectopic expression of *pal-1*. This may be a result of introducing multiple copies of *pal-1* in the transformation experiments, or the removal of a *cis*-acting control element.

We cloned the *pal-1* gene on the basis of marker rescue. We isolated a 5.5-kilobase (kb) genomic fragment that fully rescued the *Pal-1* phenotype in DNA transformation experiments (see Fig. 2 legend). Using a genomic fragment as a probe, we isolated two independent complementary DNA clones (1.3 kb). The DNA sequence of one cDNA is shown in Fig. 2. This sequence contains one large open reading frame that is predicted to encode a 208-amino-acid protein containing a homeodomain, a protein domain found in many transcriptional regulators (Fig. 2). Therefore, *pal-1* is likely to act by binding DNA and regulating transcription.

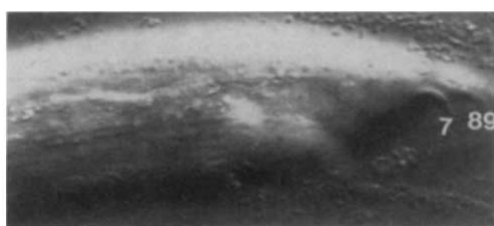
The activity of *pal-1*⁺ seems to initiate one developmental pathway (ray lineages) by somehow preventing signals from T from initiating an alternative pathway (alae lineages). Two ways in which the *pal-1* transcription factor might accomplish this seem plausible. First, *pal-1* might act directly to inhibit signal transduction; for example, it could repress transcription of a gene encoding any component of the signal transduction pathway. The failure to transduce the signal that induces alae lineages could allow the cells to initiate an alternative pathway leading to ray production. Alternatively, *pal-1* could promote the ray fate more directly in a way that precludes cell signals from having any effect. For example, *pal-1* could directly activate downstream genes that initiate ray lineages or it could repress genes that promote alae lineages. These models are illustrated in Fig. 4.

The genes *mab-5* and *lin-22* are potential downstream targets

a Wild type



b *pal-1*⁻



c *pal-1*⁻ *mab-5* (gf)

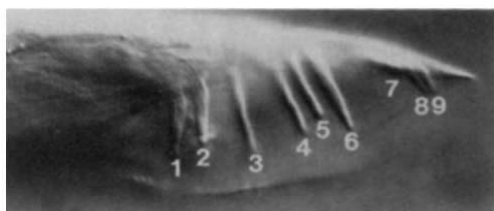


FIG. 3 A *mab-5* gain-of-function mutation overcomes the requirement for *pal-1* activity in V6. a, Wild-type male tail, showing the nine sensory rays present on each side of the animal. b, Tail of a *pal-1*(e2091); *him-5*(e1490) male. Only the T-derived rays 7, 8 and 9 are present in the tail. V6 produces alae instead of rays 2–6 in *pal-1*(e2091) mutants. The V5-derived ray 1 is not in the tail because the ray 1 neuroblast does not migrate past the transformed V6 descendants into the tail¹. c, Tail of a *pal-1* *mab-5*(gf) male (*unc-79*(e1068) *pal-1*(e2091) *mab-5*(e1751); *him-5*(e1490)). In these animals, the suppression of *pal-1* by *e1751* is virtually complete. On only 5 out of 100 sides were there fewer than the wild-type number of rays, and in those cases V6 made a few rays. By comparison, in *pal-1*(e2091) single mutants, 85% of the sides have no V6-derived rays¹.

of the cell signals and of *pal-1*. The *mab-5* gene is a homeotic gene that contains an *Antennapedia*-like homeobox and allows V6 as well as many other epidermal, neural and mesodermal cells in the posterior body region to generate the rays and other structures that characterize that region^{4,6}. The gene *lin-22* is necessary for anterior V cells to generate alae rather than ray lineages⁷. On the basis of genetic data, we previously proposed that *mab-5* and *lin-22* function in a genetic switch to control whether alae or ray lineages develop. In the model, in anterior cells, signals from neighbours enhance *lin-22* activity, which in turn inhibits *mab-5* activity and initiates alae lineages. In V6 (or its descendants), *pal-1* activity blocks or overrides the signalling pathway, and this in turn allows *mab-5* activity to inhibit *lin-22* and initiate ray lineages¹.

The idea that *pal-1* activates *mab-5* in the V6 cell is supported by properties of a *mab-5* gain-of-function mutation, *e1751* (ref. 8), which causes transformations opposite to those of *mab-5* null alleles in several tissues. This mutation alters DNA sequences upstream of *mab-5*, and so may lead to inappropriate *mab-5* transcription (S. Salser and C.K., unpublished results). To find out whether this altered *mab-5* gene could initiate ray lineages without *pal-1*⁺ activity, we constructed the *pal-1*(e2091) *mab-5*(e1751) double mutant. The *pal-1* V6 mutant phenotype is suppressed (Fig. 3). The fact that the *mab-5* gain-of-function

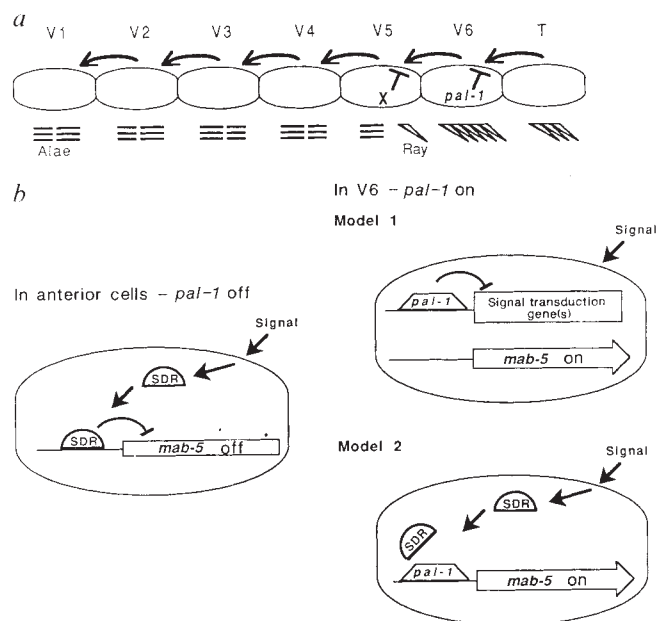


FIG. 4 Genetic and molecular models for the action of cell signals and the gene *pal-1* in establishing the alae/ray pattern. a, A formal genetic model. Signals (indicated by arrows) from cells towards their neighbours initiate alae lineages. It is possible that signalling occurs between progeny of V cells, rather than the V cells themselves. The gene *pal-1* functions within V6 to prevent T from inducing alae lineages in V6, and to allow V6 to generate ray lineages. In the absence of the T cell (after ablation) the *pal-1* gene is no longer needed for V6 to generate ray lineages¹. Because the V5 cell is not affected by our *pal-1* mutants, we have proposed that a *pal-1*-like activity acts in a similar manner to allow the production of this V5-derived ray. Other factors must be responsible for additional differences between the ray lineages; for example, the number of rays produced by different cells. b, Two ways in which *pal-1* could function as a transcriptional regulator to change the responsiveness of V6 to intercellular signals. In both models, we suggest that a signal-dependent repressor (SDR) represses transcription of *mab-5* in anterior cells, after being activated in response to intercellular signals by some mechanism. In model 1, wild-type *pal-1* blocks the response to cell signals in V6 (or its descendants) by repressing a gene encoding a component of the signal transduction pathway. In model 2, *pal-1* acts directly on a downstream gene (such as *mab-5* or possibly *lin-22*) to neutralize or override the effects of intercellular signals. Here we show *pal-1* activating *mab-5* transcription by preventing SDR from repressing *mab-5*. The *pal-1* gene could simply prevent SDR from binding, or it might act as a transcriptional activator.

mutation overcomes the requirement for *pal-1*⁺ function suggests that, in the wild type, the ultimate effect of *pal-1* is to activate *mab-5*. Thus *mab-5* is a candidate for a downstream gene directly regulated by *pal-1*. One way in which *pal-1* could activate a downstream gene such as *mab-5* would be for the *pal-1* protein to bind to the gene in such a way that a second transcriptional regulator—one that is dependent on signals from T—could not function. This would explain why *pal-1* function is required only if the signalling cell T is present (see Fig. 4).

Whatever the means by which *pal-1* activity prevents V6 from responding to intercellular signals, these findings indicate that a mechanism exists for creating unequal signal responsiveness between precursor cells that otherwise have similar developmental potentials. Because this differential responsiveness plays such an important part in producing the pattern of alae and rays, it will be interesting to discover how *pal-1* function can be targeted so precisely to V6. □

Received 3 January; accepted 28 February 1991.

1. Waring, D. A. & Kenyon, C. *Cell* **60**, 123–131 (1990).
2. Sulston, J. & Horvitz, H. R. *Dev. Biol.* **56**, 110–156 (1977).
3. Sulston, J. & White, J. *Dev. Biol.* **78**, 577–597 (1980).
4. Kenyon, C. *Cell* **46**, 477–487 (1986).
5. Austin, J. & Kimble, J. *Cell* **51**, 589–599 (1987).
6. Costa, M., Weir, M., Coulson, A., Sulston, J. & Kenyon, C. *Cell* **55**, 747–756 (1988).
7. Horvitz, H., Sternberg, P., Greenwald, I., Fixsen, W. & Ellis, H. *Cold Spring Harb. Symp. quant. Biol.* **48**, 453–463 (1983).
8. Hedgecock, E., Culotti, J., Hall, D. & Stern, B. *Development* **100**, 365–382 (1987).
9. Sulston, J., Schierenberg, E., White, J. & Thomson, J. *Dev. Biol.* **100**, 64–119 (1983).
10. Rosenbluth, R., Cuddeford, C. & Baillie, D. *Genetics* **109**, 493–511 (1985).

11. Burglin, T. R., Finney, M., Coulson, A. & Ruvkun, G. *Nature* **341**, 239–243 (1989).
12. Scott, M. P. & Carroll, S. B. *Cell* **51**, 689–698 (1987).
13. Fire, A. *EMBO J.* **5**, 2673–2680 (1986).
14. Krause, M. & Hirsh, D. *Cell* **49**, 753–761 (1987).

ACKNOWLEDGEMENTS. We thank H. Madhani for his contribution to the mapping of *pal-1*; members of C.K.'s laboratory and C. Goutte for discussions and comments on the manuscript; C. Martin for the cDNA library; J. Sulston and A. Coulson for cosmid clones; and E. Hedgecock, D. Thierry-Mieg, R. Hoskins and A. Chisom for sharing unpublished information about *mab-5(e1751)*. Some nematode strains were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources. This work was supported by the NIH. D.W. was supported by the NIH and the Lucille P. Markey Foundation.

Phosphorylation of two small GTP-binding proteins of the Rab family by p34^{cdc2}

Eric Bailly*, Mary McCaffrey†, Nicolas Touchot‡§, Ahmed Zahraoui‡, Bruno Goud† & Michel Bornens*

* Centre de Génétique Moléculaire, CNRS, 91198 Gif-sur-Yvette, France

† Unité de Génétique Somatique (URA CNRS 361), Institut Pasteur, 75724 Paris, France

‡ Faculté de médecine Lariboisière-Saint Louis, INSERM unité 248, 75010 Paris, France

§ Present address: Department of Biology, Lewis Thomas Laboratory, Princeton University, Princeton, New Jersey 08544, USA

ENTRY of a cell into mitosis induces a series of structural and functional changes including arrest of intracellular transport^{1–4}. Knowledge of how the mitotic cycle is driven progressed substantially with the identification of the p34^{cdc2} protein kinase as a subunit of maturation-promoting factor^{5–7}, the universal regulating component of the mitotic cycle⁸. Activation of the kinase at the onset of mitosis⁹ is thought to trigger the important mitotic events by phosphorylating key proteins¹⁰. Small guanine nucleo-

tide-binding proteins have been implicated in regulating transport pathways. For instance, two small Ras-related GTP-binding proteins, Sec4p and Ypt1p, control distinct stages of the secretory pathway in budding yeast^{11–15}. The GTP-binding proteins of the Rab family in rats and humans^{16,17} display strong homologies with Sec4p and Ypt1p, and might therefore also be involved in regulating intracellular transport. Indeed, distinct Rab proteins are located in the exocytotic and endocytotic compartments^{18–21}. Interruption of vesicular transport during mitosis might involve modification of these proteins. We now present biochemical evidence for a mitosis-specific p34^{cdc2} phosphorylation of Rab1Ap and Rab4p. By contrast, Rab2p and Rab6p are not phosphorylated. We also show that the distribution of Rab1Ap and Rab4p between cytosolic and membrane-bound forms is different in interphase and mitotic cells. This may provide a clue to the mechanism by which phosphorylation could affect membrane traffic during mitosis.

To determine whether the expression of Rab proteins varies during the cell cycle, whole cell extracts from interphase and mitotic HeLa cells were probed with specific polyclonal rabbit antisera raised against human Rab proteins expressed in *Escherichia coli*. The levels of Rab1A, Rab2, Rab4 and Rab6 proteins did not seem to change significantly as cells proceeded from interphase to mitosis (Fig. 1a), ruling out any major regulation of the expression of these proteins during the cell cycle. Because the onset of mitosis is known to coincide with a

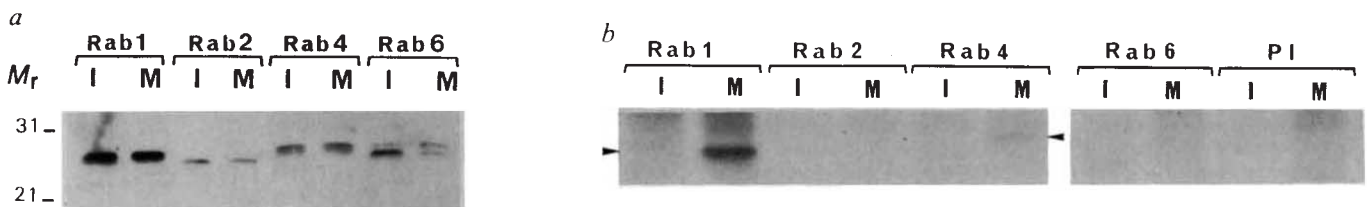


FIG. 1 Immunoblot and ³²P-immunoprecipitation analysis of human Rab1A, Rab2, Rab4 and Rab6 proteins during the cell cycle. Cell extracts were prepared from either interphasic (I) or mitotic (M) synchronized HeLa cells. **a**, Comparison of Rab1Ap, Rab2p, Rab4p and Rab6p content in interphasic and mitotic whole HeLa cell extracts. Total proteins (30 µg) were loaded in each lane and probed by western blotting with the corresponding rabbit immune sera. **b**, Immunoprecipitation of Rab proteins from *in vivo* ³²P-labelled cell lysates. Preimmune serum (PI) of the rabbit that produced anti-Rab6p was used as a blank. Immunoprecipitates were visualized by autoradiography of SDS gels. Electrophoretic mobilities of bacterially expressed Rab1Ap and Rab4p are indicated by arrowheads.

METHODS. Interphasic HeLa cells were treated with 2.5 mM thymidine for 16 h and then recovered. Cells accumulated in mitosis by a 16-h treatment with 1 µM Nocodazole were collected by pipetting. Characterization of the anti-Rab antibodies has been described elsewhere^{18,32}. **a**, Samples were

prepared and detected as previously described³³ except that the secondary peroxidase-conjugated antibody was detected with the ECL system according to the manufacturer's instructions (Amersham). **b**, Synchronized cells were incubated for 90 min in phosphate-free RPMI medium containing 7% dialysed FCS before addition of 350 µCi ml⁻¹ ³²P_i (Amersham). Cells were labelled for 3 h in the continuous presence of the same concentration of thymidine or Nocodazole, and were lysed in modified RIPA buffer (0.1 M Tris-HCl, pH 8.3, 2 mM EDTA, 0.5% Nonidet P-40, 0.5% deoxycholate, 0.1% SDS, 150 mM NaCl, 10 mM NaF, 50 µM sodium orthovanadate, 20 mM β-glycerophosphate, 5 µg ml⁻¹ leupeptin/pepstatin, 50 µg ml⁻¹ aprotinin and 1 mM PMSF). Immunoprecipitation was performed as described previously³³ and analysed on 12.5% polyacrylamide-SDS gels. After staining with Coomassie blue, gels were dried and exposed to Kodak XA-S films for 72 h at -80 °C with an intensifying screen (Dupont Cronex).

burst of phosphorylation, we then investigated whether these proteins were phosphorylated in M phase. Rab proteins were immunoprecipitated from *in vivo* ^{32}P -labelled interphasic and mitotic cell lysates and analysed by SDS-PAGE and autoradiography (Fig. 1b). Whereas Rab2p and Rab6p were not phosphorylated in either interphase or mitosis, ^{32}P -labelled bands comigrating with bacterially produced Rab1Ap and Rab4p were immunoprecipitated from mitotic cell lysates only. The difference in signal intensity between Rab1Ap and Rab4p was in good agreement with the comparative amounts of these proteins as judged by immunodetection (see Fig. 1a). These results indicate that Rab1Ap and Rab4p, but not Rab2p and Rab6p, are specifically phosphorylated during mitosis.

We next attempted to identify the protein kinase responsible for the M phase-specific phosphorylation. Examination of the deduced amino-acid sequences¹⁷ showed that only Rab1Ap and Rab4p contained putative consensus motifs for a p34^{cdc2} phosphorylation site in their C-terminal region (Fig. 2a). To test whether p34^{cdc2} was the relevant protein kinase, purified Rab1Ap and Rab4p were used as substrates in an *in vitro* phosphorylation assay with immunoprecipitated p34^{cdc2} from either interphase or mitotic cell extracts (Fig. 2b). Histone H1 and Rab6p were included as positive and negative controls, respectively. Accordingly, mitotic p34^{cdc2} which displayed a high

histone H1 kinase activity was also found to phosphorylate purified Rab1Ap as well as Rab4p but not Rab6p. The prominent phosphorylation of histone H1 compared with that of Rab proteins was expected, as the latter contain one putative site for p34^{cdc2} phosphorylation whereas histone H1 is phosphorylated on multiple sites²². Conversely, interphase p34^{cdc2} displayed a very low kinase activity for all exogenous substrates. Similar experiments with the catalytic subunit of the cyclic AMP-dependent type-II protein kinase revealed that none of the Rab proteins tested could be phosphorylated *in vitro* (data not shown).

To obtain further evidence for involvement of p34^{cdc2} in Rab1Ap and Rab4p phosphorylation, we determined whether p34^{cdc2}-depleted mitotic cell extracts could still phosphorylate exogenously added purified Rab proteins. We used the high

a

Hrab1 (184-195)	N V K I Q S T P V K - - - - -	Q S
Hrab2 (180-199)	G I K I G P Q H A A T N A T H A G N	Q G
Hrab4 (180-199)	G I Q Y G D A A L R Q L R S P R R T	Q A
Hrab6 (191-200)	I K L E K P Q E - - - - -	Q P

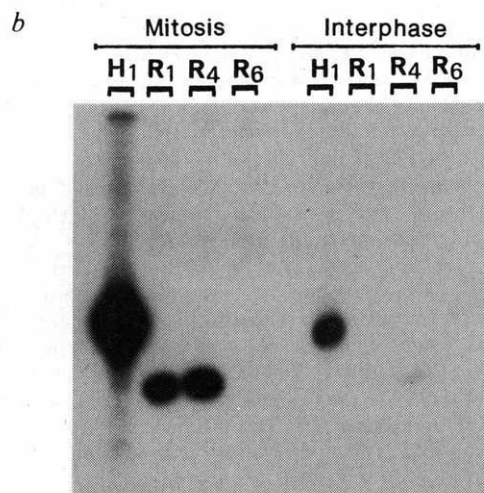


FIG. 2 Rab1Ap and Rab4p are phosphorylated *in vitro* by p34^{cdc2}. a, Comparison of the C-terminal amino-acid sequences (single-letter notation) of human Rab1A, Rab2, Rab4 and Rab6 proteins. Localization of p34^{cdc2} consensus sequences in Rab1Ap and Rab4p is indicated by bold letters. Putative phosphorylated residues are marked with an asterisk. Sequences were aligned according to Zahraoui *et al.*¹⁷. b, *In vitro* p34^{cdc2} phosphorylation. Immunoprecipitates of p34^{cdc2} from mitotic and interphase cell lysates were assayed for *in vitro* phosphorylation activity with purified histone H1 (H1), Rab1Ap (R1), Rab4p (R4) and Rab6p (R6) as substrates. Incorporation of radioactive phosphate was followed by SDS-PAGE and autoradiography of the gels.

METHODS. Synchronized HeLa cells were obtained as described in the legend to Fig. 1. Immunoprecipitation of p34^{cdc2} was carried out according to Draetta and Beach⁹ with a polyclonal antibody raised against the octapeptide LDNQIKKM corresponding to the C terminus of mammalian-specific cdc2 homologues. *In vitro* phosphorylation of histone H1 (3.3 µg, Boehringer Mannheim) or purified Rab proteins (5 µg) was performed in a 15-µl final volume of TKM buffer (50 mM Tris-HCl, pH 7.4, 25 mM KCl, 5 mM MgCl₂ and 0.5 mM DTT) containing 100 µM cold ATP and 1.5 µCi [γ - ^{32}P]ATP (Amersham, 5,000 Ci mmol⁻¹). After a 15-min incubation at 37 °C the reaction was stopped by addition of 15 µl of 2 × SDS sample buffer. Aliquots (5 µl) were electrophoresed on 12% SDS-PAGE minigels. After staining with Coomassie blue, gels were dried and autoradiographed at -80 °C for 24 h with an intensifying screen.

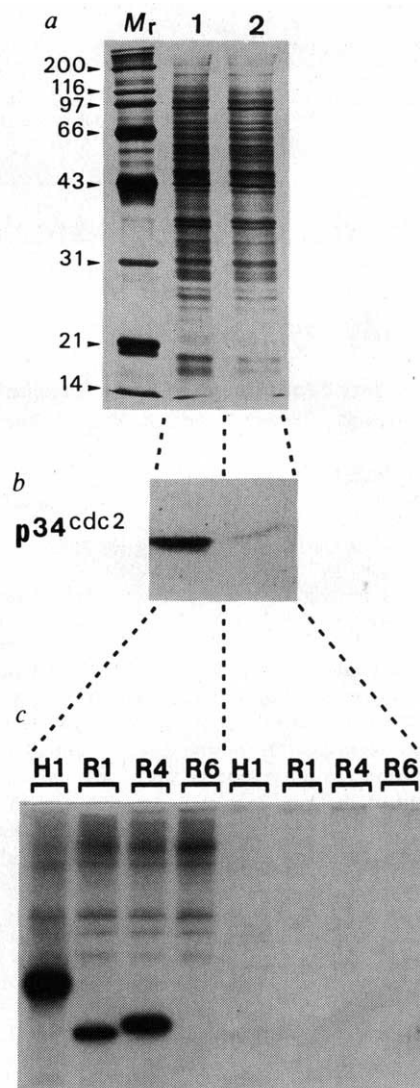


FIG. 3 Depletion of mitotic extracts of p34^{cdc2} abolishes phosphorylation of Rab1Ap and Rab4p. Aliquots of mitotic extracts preincubated either with control Sepharose beads (lane 1) or with p13^{suc1}-carrying Sepharose beads (lane 2) were analysed in the following way. a, Protein patterns of both extracts after electrophoresis on a 10% polyacrylamide-SDS gel and silver-staining; M_r, relative molecular mass. b, Amount of p34^{cdc2} in both extracts as revealed by western blotting. c, Kinase activity of both extracts on exogenous substrates: H1, histone H1; R1, Rab1Ap; R4, Rab4p and R6, Rab6p. The gel was exposed for 24 h at -80 °C with an intensifying screen.

METHODS. The same mitotic extracts as described in Fig. 1 were depleted of endogenous p34^{cdc2} by two successive incubations with p13^{suc1} beads prepared as detailed elsewhere²⁹. Bead supernatants were then analysed either by SDS-PAGE electrophoresis followed by silver-staining of the gel or by immunodetection of p34^{cdc2} with the PSTAIR antibody³². The phosphorylation assay was carried out as described in the legend to Fig. 2.

affinity of p34^{cdc2} for the fission yeast *suc1*⁺ gene product, p13^{suc1} (ref. 5). After incubation with p13^{suc1} or control Sepharose beads, the cell lysate was recovered and tested for both p34^{cdc2} content and kinase activity. Depletion of p34^{cdc2} in mitotic cell

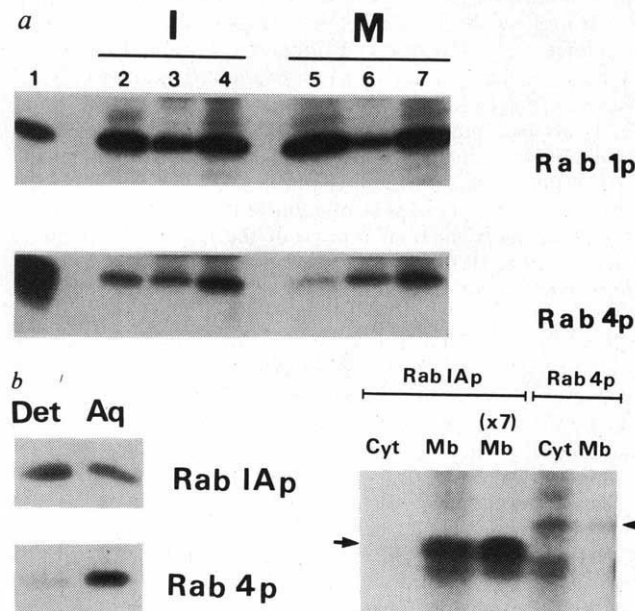


FIG. 4 Partitioning of Rab1Ap and Rab4p. *a*, Immunoblot analysis and measurement of membrane-bound and cytosolic forms of Rab1Ap and Rab4p in interphasic and mitotic cells. The following samples were analysed by immunoblotting with anti-Rab1Ap and anti-Rab4p antibodies: 2, 3, 4, extracts from interphasic cells (I), 5, 6, 7, extracts from mitotic cells (M). Lanes 4 and 7, post-nuclear supernatants; lanes 3 and 6, high-speed supernatants (cytosol); lanes 2 and 5, high-speed pellets (membranes). In lane 1, the samples were 50 ng purified Rab1Ap or Rab4p. The autoradiograph was exposed for 3.5 h at -80°C . The percentage of total cellular Rab1Ap and Rab4p present in the cytosol of interphase and mitotic extracts was:

	Interphase	Mitosis
Rab1Ap	31 $\pm$ 6	19 $\pm$ 4
Rab4p	48 $\pm$ 5	74 $\pm$ 3

The indicated values are the mean $\pm$ s.d. of two to four independent experiments. *b*, Immunoblot analysis of Rab1Ap and Rab4p in mitotic cytosol extracted with Triton X-114. A high-speed supernatant prepared as described above (lane 6, *a*) from mitotic cells was extracted twice with Triton X-114 (ref. 34) and the resulting detergent (Det) and aqueous (Aq) phases were analysed by immunoblotting. The autoradiograph was exposed for 14 h at -80°C . *c*, Subcellular distribution of phosphorylated Rab1Ap and Rab4p in mitotic cells. Cytosolic (Cyt) and membrane (Mb) fractions, prepared as described above, from *in vivo* ^{32}P -labelled mitotic cells were used for immunoprecipitation experiments. Immunoprecipitates were made visible by autoradiography of SDS gels. Lane ($\times 7$) Mb corresponds to immunoprecipitation of seven times the volume assayed in lane Mb. Arrows on the left and right hand of the figure indicate the position of Rab1Ap and Rab4p, respectively. The autoradiograph was exposed for 96 h at -80°C . **METHODS.** Interphase and mitotic HeLa cells were obtained as described in the legend to Fig. 1. *a*, Post-nuclear supernatants (PNS) were prepared as previously described¹⁸ and centrifuged at 150,000*g* for 1 h at 4°C to generate cytosol and crude membrane fractions. Aliquots of the different fractions were separated on 12% polyacrylamide-SDS gels. The amount of protein loaded onto the gels was standardized to take into account the distribution of total proteins (PNS) between cytosolic and membrane fractions. Immunoblotting and quantitation using ^{125}I -labelled protein A (Amersham) was then performed as previously described¹⁸. *b*, Cytosol (200 μl) prepared as described above was extracted twice with an equal volume of 10 mM Tris, pH 7.4, 0.05 M sucrose, 0.3 M NaCl, 2% Triton X-114, protease inhibitors (leupeptin, antipain, pepstatin and aprotinin) to 0.1 $\mu\text{g ml}^{-1}$ and PMSF to 1 mM. The resulting aqueous and detergent phases were precipitated with TCA and analysed by immunoblotting. *c*, *In vivo* labelling of mitotic HeLa cells with ^{32}P , and immunoprecipitation were carried out as described in the legend to Fig. 1. *b*, Cytosolic and membrane fractions were prepared as described above and their volume normalized to reflect directly the proportion of Rab proteins in each fraction.

lysates abolished histone H1 as well as Rab1Ap and Rab4p phosphorylation (Fig. 3). In agreement with the results presented in Fig. 2*b*, both control and p34^{cdc2}-depleted interphase extracts exhibited very low kinase activities for exogenous Rab proteins (data not shown). These *in vitro* experiments clearly indicate that p34^{cdc2} is the mitotic protein kinase that phosphorylates Rab1Ap and Rab4p.

So how might phosphorylation affect the function of Rab1Ap and Rab4p? Results indicate that the GTP-binding abilities of both proteins are not modified after p34^{cdc2} phosphorylation (data not shown). For several small GTP-binding proteins, the C-terminal region is crucial for membrane attachment, an event required for their proper functioning^{23–25}. That the consensus phosphorylation sequence resides in the C-terminal region (Fig. 2*a*) prompted us to compare the partitioning of Rab1Ap and Rab4p on membranes and in cytosolic fractions between interphase and mitotic cells. Most of the Rab4p was cytosolic in mitotic cells as opposed to interphase cells where slightly more than half of the Rab4p present partitioned in the membrane fraction. In contrast, a slight but reproducible reduction of Rab1Ap was detected in the mitotic cytosol (Fig. 4*a*). We were aware of the possibility of artefactual observations in partitioning, due to membrane fragmentation during mitosis. Our results argue against this possibility, however, as the Rab1Ap and Rab4p partitioning trends are in opposing senses in mitotic extracts. Also, the partitioning of Rab2p and Rab6p does not change significantly between interphase and mitosis (data not shown). To address this question more directly, the cytosol from mitotic cells was extracted with Triton X-114. Most of the cytosolic form of Rab4p partitioned into the aqueous phase, again strongly arguing against membrane contamination (Fig. 4*b*). For Rab1Ap, a significant amount of cytosolic protein partitioned in the detergent phase, an observation which has a precedent in the case of Rab3p (ref. 26). To uncover further evidence for the relevance of Rab4p phosphorylation to its partitioning in mitosis we studied the distribution of the mitotically phosphorylated protein between membranes and cytosol. As expected, most of the phosphorylated Rab1Ap and Rab4p were found in the membrane and cytosolic fractions, respectively (Fig. 4*c*).

Our study establishes that both Rab1Ap and Rab4p are phosphorylated *in vitro* and *in vivo* by the M phase-inducing protein kinase p34^{cdc2}, a result which identifies these membrane-associated proteins as a new class of p34^{cdc2} substrates. Rab1Ap, the mammalian homologue²⁷ of yeast Ypt1p, probably regulates a transport event between endoplasmic reticulum and the Golgi apparatus and/or the Golgi compartments^{15,28}. As in the case of Ypt1p in *Saccharomyces cerevisiae* (ref. 16), Rab1Ap is located in the Golgi region of the cell (N.T. and B.G., unpublished data). The precise localization of Rab4p is presently unknown but recent data indicates that it is associated with early endosomes²⁹. Tuomikoski *et al.*³⁰ have reported that active p34^{cdc2} severely inhibits *in vitro* fusion of endocytotic vesicles. Taken together, these observations suggest that phosphorylation of Rab4p by p34^{cdc2} kinase may be fundamental to the mechanism by which endocytosis is arrested in mitotic cells.

Studies on Sec4p indicate that this protein moves from post-Golgi vesicles to plasma membrane in concert with vesicle transport and then returns to vesicles by way of the cytoplasm^{12,23}. On the basis of these results, it has been postulated that vesicular transport is coupled to the cycle of small GTP-binding proteins between a cytosolic and a membrane-bound form^{23,31}. Our data indicate that p34^{cdc2} phosphorylation may alter the cyclic process for Rab1Ap and Rab4p by perturbing their partitioning and lead to the arrest of intracellular transport during mitosis. □

Received 3 December 1990; accepted 7 March 1991.

- Warren, G., Featherstone, C., Griffiths, G. & Burke, B. *J. Cell Biol.* **97**, 1623–1628 (1983).
- Featherstone, C., Griffiths, G. & Warren, G. *J. Cell Biol.* **101**, 2036–2046 (1985).

3. Warren, G. *Trends biochem. Sci.* **10**, 439–443 (1985).
4. Warren, G. *Nature* **342**, 857–858 (1989).
5. Dunphy, W. G., Brizuela, L., Beach, D. & Newport, J. *Cell* **54**, 423–431 (1988).
6. Gautier, J., Norbury, C., Lohka, M., Nurse, P. & Maller, J. *Cell* **54**, 433–439 (1988).
7. Labbé, J.-C., Lee, M., Nurse, P., Picard, A. & Dorée, M. *Nature* **335**, 251–254 (1988).
8. Nurse, P. *Nature* **344**, 503–508 (1990).
9. Draetta, J. & Beach, D. *Cell* **54**, 17–26 (1988).
10. Moreno, S. & Nurse, P. *Cell* **61**, 549–551 (1990).
11. Salminen, A. & Novick, P. *Cell* **49**, 527–538 (1987).
12. Goud, B., Salminen, A., Walworth, N. & Novick, P. *Cell* **53**, 753–768 (1988).
13. Schmitt, H. D., Wagner, P., Pfaff, E. & Gallwitz, D. *Cell* **47**, 401–412 (1986).
14. Segev, N., Mulholland, J. & Bostein, D. *Cell* **52**, 915–924 (1988).
15. Bacon, R. A., Salminen, A., Ruohola, H., Novick, P. & Ferro-Novick, S. *J. Cell Biol.* **109**, 1015–1022 (1989).
16. Touchot, N., Chardin, P. & Tavittian, A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 8210–8214 (1987).
17. Zahraoui, A., Touchot, N., Chardin, P. & Tavittian, A. *J. biol. Chem.* **264**, 12394–12401 (1989).
18. Goud, B., Zahraoui, A., Tavittian, A. & Saraste, J. *Nature* **345**, 553–556 (1990).
19. Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K. & Zerial, M. *Cell* **62**, 317–329 (1990).
20. Mollard, G. F. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1988–1992 (1990).
21. Darchen, F. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 5692–5696 (1990).
22. Langan, T. A. *J. biol. Chem.* **257**, 14835–14846 (1982).
23. Walworth, N., Goud, B., Kabenell, A. & Novick, P. *EMBO J.* **8**, 1685–1693 (1989).
24. Molenaar, C. M. T., Prange, R. & Gallwitz, D. *EMBO J.* **7**, 1167–1177 (1988).
25. Willumsen, B. E., Christensen, A., Hubbert, N. L., Papageorge, A. G. & Lowy, D. R. *Nature* **310**, 583–586 (1984).
26. Fischer, v. Mollard, G. et al. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1988–1992 (1990).
27. Haubruck, H., Prange, R., Vorgias, C. & Gallwitz, D. *EMBO J.* **8**, 1427–1432 (1989).
28. Baker, D., Wuestehube, L., Scheckman, R., Bostein, D. & Segev, N. *Proc. natn. Acad. Sci. U.S.A.* **87**, 355–359 (1990).
29. Van der Sluijs, P. et al. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
30. Tuomikoski, T., Felix, M. A., Dorée, M. & Gruenberg, J. *Nature* **342**, 942–945 (1989).
31. Bourne, H. *Cell* **53**, 669–671 (1988).
32. Maridonneau-Parini, I., Yang, C. Z., Bornens, M. & Goud, B. *J. clin. Invest.* **87**, 901–907 (1991).
33. Bailly, E., Dorée, M., Nurse, P. & Bornens, M. *EMBO J.* **8**, 3985–3995 (1989).
34. Bordier, C. *J. biol. Chem.* **256**, 1604–1607 (1981).

ACKNOWLEDGEMENTS. We thank J. Hayles, C. Norbury and P. Nurse for the gift of the anti-p34^{cdc2} antibody and M. Dorée and J.-C. Cavadore for providing us with p13^{suc2} beads and PSTAIR antibody. We also thank G. Buttin for support and R. M. Rios Sanchez for helpful discussions. I. Gaspard and D. Meur for artwork and S. Brown for proofreading. This work was supported by the CNRS, the Université Pierre et Marie Curie, the Ligue Nationale Française Contre le Cancer (E.B.) and the European Community (M.M.C.).

Inhibition of telomerase by G-quartet DNA structures

Alan M. Zahler*, James R. Williamson*,
Thomas R. Cech & David M. Prescott

Department of Molecular, Cellular and Developmental Biology,
Department of Chemistry and Biochemistry,
Howard Hughes Medical Institute, University of Colorado,
Boulder, Colorado 80309-0215, USA

THE ends or telomeres of the linear chromosomes of eukaryotes are composed of tandem repeats of short DNA sequences, one strand being rich in guanine (G strand) and the complementary strand in cytosine^{1,2}. Telomere synthesis involves the addition of telomeric repeats to the G strand by telomere terminal transferase (telomerase)^{3–6}. Telomeric G-strand DNAs from a variety of organisms adopt compact structures⁷, the most stable of which is explained by the formation of G-quartets^{8,9}. Here we investigate the capacity of the different folded forms of telomeric DNA to serve as primers for the *Oxytricha nova* telomerase *in vitro*. Formation of the K⁺-stabilized G-quartet structure in a primer inhibits its use by telomerase. Furthermore, the octanucleotide T₄G₄, which does not fold, is a better primer than (T₄G₄)₂, which can form a foldback structure^{7–10}. We conclude that telomerase does not require any folding of its DNA primer. Folding of telomeric DNA into G-quartet structures seems to influence the extent of telomere elongation *in vitro* and might therefore act as a negative regulator of elongation *in vivo*.

The role of telomerase in telomere extension *in vivo* has been established in *Tetrahymena*¹¹. Analogous activity in *Oxytricha*, *Euplotes* and human cells suggests that the function of this enzyme in telomere replication is general^{12–14}. In this study we

use the species *Oxytricha nova*, whose unusually large number of macronuclear DNA molecules (~10⁸ per cell) has facilitated studies of telomeres and telomerase¹⁵.

The G-quartet is a square planar, hydrogen-bonded array of four guanine bases, each of which is both the donor and acceptor of a Hoogsteen base pair¹⁶. For a telomeric G strand to be intramolecularly folded into a G-quartet, it should have four telomeric repeats^{8,9}. The structure is stabilized by physiological concentrations of K⁺ and, to a lesser extent, Na⁺ (refs 8, 9 and 17). Two-repeat molecules such as (T₄G₄)₂ also form compact intramolecular structures^{7–10} but do not show cation-specific stabilization characteristic of G-quartets⁸.

We synthesized a series of oligonucleotides oxy-*n* (*n* = 1–8), where *n* is the number of repeats of the *Oxytricha* telomeric DNA sequence d(TTTTGGGG). The mobility of these oligonucleotides in gels containing Tris-borate buffer depends monotonically on their length (Fig. 1), but when 20 mM Na⁺ or 20 mM K⁺ is included there is a discontinuity in the relative mobilities of oxy-3 and oxy-4, indicating that oxy-4 is more compact than oxy-3. Oxy-8 also forms a species more compact than oxy-7. These data show that each Na⁺- or K⁺-induced intramolecular folding event requires four repeats of single-stranded telomeric DNA.

The activity of the *Oxytricha* telomerase was tested using these same oligonucleotides as primers in the presence and absence of monovalent cations. All reactions contained 10 mM MgCl₂ because Mg²⁺ is an absolute requirement for telomerase. The efficiency of extension of each oligonucleotide primer was comparable in the absence of added monovalent cations and in the presence of 20 mM Na⁺ (Fig. 2a), but when 20 mM K⁺ was included, oxy-4 was far less effective as a primer (lane K of set 4). The activity of oxy-5 and oxy-8 was markedly reduced in K⁺, although not as much as that of oxy-4; oxy-6 and oxy-7 activities were only slightly reduced.

There is no inhibition by 20 mM Na⁺ or K⁺ on telomerase itself, as the activity of primers with *n* < 4 was unaffected by the presence of these ions (Fig. 2a). Contamination of primer samples or the possible inhibition of telomerase by the longer primers was discounted by experiments in which primers were mixed together (data not shown). In similar experiments using oligonucleotides composed of telomeric repeats from *Tetrahymena*, extension of tet-2 (d(TTGGGG)₂) was not affected by monovalent cations, whereas extension of tet-4 (d(TTGGGG)₄) was reduced compared with tet-2, and greatly reduced in the presence of K⁺ (data not shown).

Oxytricha telomerase is somewhat processive *in vitro*¹⁸, so K⁺ could affect primer extension at either the initiation or the elongation step. Omitting dGTP from the reaction causes elongation to terminate after addition of the first four T residues. Addition of K⁺ reduced utilization of oxy-4 but not of oxy-2 (Fig. 2b). Therefore the effect of the structure of a primer on its use by telomerase can be attributed at least in part to the initiation step.

When two different substrates at equal concentrations compete for the same enzyme, their relative rates of reaction are determined by their *k*_{cat}/*K*_m values; this conclusion holds at all substrate concentrations¹⁹. An equimolar mixture of oxy-1 and oxy-2 was incubated with telomerase using TTP as the only nucleotide. Oxy-1 reacted about twice as fast as oxy-2 whether or not Na⁺ or K⁺ was present (data not shown). Thus oxy-1 has a slightly greater value of *k*_{cat}/*K*_m, the rate constant that measures the reaction of free enzyme and free substrate. Because oxy-1 does not form intramolecular structures under these conditions and forms a small amount of intermolecular structure only in K⁺ (Fig. 1), but was equally active as a primer in the absence and presence of K⁺ and Na⁺, we conclude that unstructured primers can be efficiently recognized and extended by telomerase. Confirmation and elaboration of this limited kinetic analysis awaits purification of the enzyme.

Besides the specific monovalent cation effect on the extension

* Present addresses: Fred Hutchinson Cancer Research Center, Seattle, Washington 98104, USA (A.M.Z.); and Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (J.R.W.).

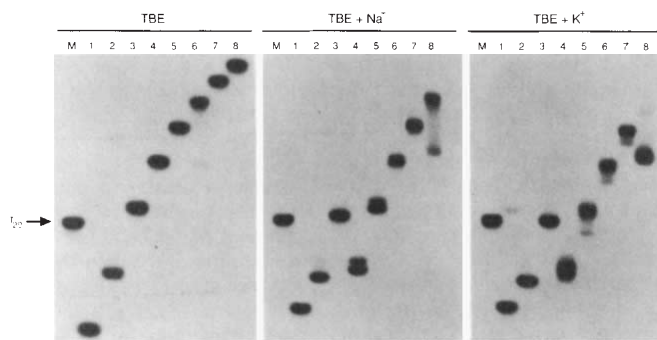


FIG. 1 Folding of *Oxytricha* telomeric oligonucleotides. Single-stranded oligonucleotides composed of $n=1-8$ repeats of the *Oxytricha* telomeric sequence $d(TTTTGGGG)_n$ (oxy-1 to oxy-8, lanes 1-8 in each panel) were subjected to gel electrophoresis under non-denaturing conditions in Tris-borate-EDTA buffer, pH 8.3 (TBE), or in TBE with added Na^+ or K^+ . The unstructured oligonucleotide dT_{20} was run as a marker (lane M). Native gels were run as described⁸, except that the temperature was maintained at 15 °C and NaCl or KCl as added to 20 mM.

of oxy-4, oxy-4 was a much poorer primer than oxy-2 even when only Mg^{2+} was present (Fig. 2a and b). Under all conditions the extension of oxy-4, 5, 6, 7 and 8 was much less than that of oxy-1, 2 and 3. As discussed later, oxy-4 folds in the presence of Mg^{2+} alone, which can explain the reduced efficacy of primers with more than three repeats.

Although telomerase has to be assayed in magnesium-containing buffers, the effect of Mg^{2+} on the structure of *Oxytricha* telomeric DNA had not been previously tested. Dimethyl sulphate treatment of oxy-4 at 15 °C in the presence of 10 mM Mg^{2+} , 10 mM Mg^{2+} with 20 mM Na^+ , and 10 mM Mg^{2+} with 20 mM K^+ , all gave methylation-protection patterns similar to that for K^+ alone⁸, namely patterns indicative of G-quartet formation (data not shown). Also, the ultraviolet-induced cross-link that is characteristic of the folded form of oxy-4 (ref. 8) was observed under all three sets of ionic conditions (data not shown). Therefore in all our telomerase assays the equilibrium

for oxy-4 is shifted toward the folded form. The K^+ forms of some telomeric DNAs are much more stable than the Na^+ forms (ref. 8; I. Tinoco Jr and C. C. Hardin, personal communication), so we investigated whether this difference in stability could explain the effect of K^+ in preventing utilization of oxy-4 by telomerase.

Thermal denaturation of oxy-4 indicates that it undergoes a cooperative unfolding transition under all of our ionic conditions (Fig. 3a). The T_m for the $\text{Mg}^{2+}/\text{K}^+$ form is 30 °C higher than for the $\text{Mg}^{2+}/\text{Na}^+$ form, and 40 °C higher than for the Mg^{2+} form. The $\text{Mg}^{2+}/\text{K}^+$ form of oxy-4 is calculated to be $\sim 8 \text{ kcal mol}^{-1}$ more stable than the $\text{Mg}^{2+}/\text{Na}^+$ form at 15 °C (Fig. 3b). The hypochromicities for the oxy-4 forms under the three sets of ionic conditions are very different (Fig. 3b), as are their absorption spectra (Fig. 3c and d). Although the dimethyl sulphate protection assay showed that the three forms are fundamentally similar, their absorption spectra and hypochromicities indicate that their folded structures differ in detail.

We conclude that the most stable folded form of telomeric DNA, which we propose is a G-quartet structure, is not an obligatory primer for telomerase, and probably cannot serve as a primer at all. Furthermore, there is apparently no secondary structure requirement for the primer, as the oxy-1 octanucleotide is an effective primer. These results do not support an earlier suggestion that telomerase recognizes a folded form of single-stranded DNA^{4,7}, although there could be differences between the *Tetrahymena* and *Oxytricha* enzymes. But they are consistent with the observation that inosine substitutions, which destabilize G-quartet formation, do not prevent telomerase activity with the *Tetrahymena* enzyme²⁰. The telomere-binding protein from *Oxytricha*, a protein that caps the DNA termini and has no known enzyme activity, also selectively binds to the unfolded form of oxy-4 rather than to the most stable folded form²¹.

A model for the action of telomerase consisting of sequential binding, polymerization, and translocation steps has been proposed by Blackburn and colleagues^{5,6}. The effect of primer structure on telomerase activity can now be incorporated into this model (Fig. 4). For primers with four or more repeats, the concentration of unfolded DNA competent to bind telomerase

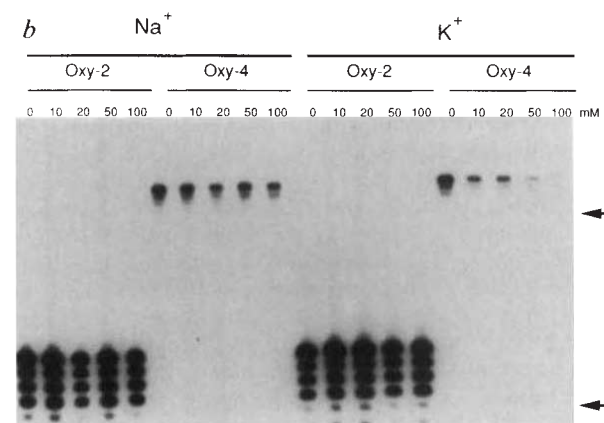
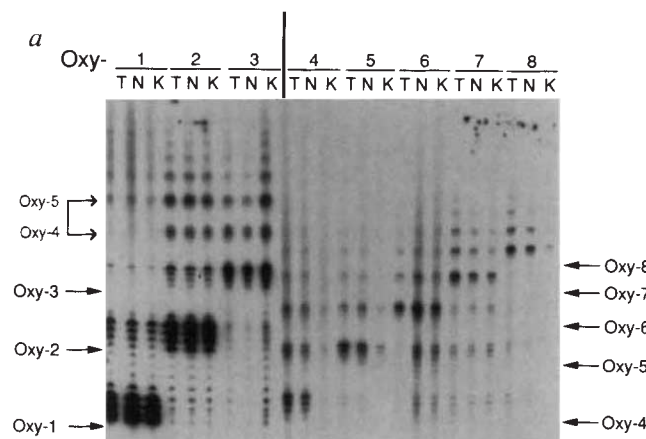


FIG. 2 Effect of Na^+ and K^+ on telomerase activity using telomeric oligonucleotides as primers. a, Standard telomerase assay on oxy-1 to oxy-8 (lane sets 1-8) in Tris-HCl and Mg^{2+} (T lanes), or with Na^+ (N lanes) or K^+ (K lanes) added to 20 mM. Only elongated products are labelled, and the approximate position of the input unlabelled DNA is indicated at the side for each oligonucleotide. b, Initiation of primer elongation by the *Oxytricha* telomerase. Omission of dGTP from the standard telomerase assay results in addition of no more than four T residues to the primer, providing a reasonable approximation to the rate of initiation of telomeric DNA synthesis. The amount of initiation with oxy-2 as primer was unaffected by the addition of Na^+ or K^+ . The amount of initiation with oxy-4 as primer was unaffected by the addition of Na^+ , but substantial reduction was caused by the addition of K^+ . Positions of input unlabelled primers are indicated at the side. METHODS. *Oxytricha nova* was cultivated and its macronuclei isolated as

described¹². Purified macronuclear pellets were suspended in 10 mM Tris-HCl, pH 8.0, 10 mM MgCl_2 , 8% w/v sucrose at a concentration of 4×10^6 macronuclei per ml and stored in aliquots at -70°C . Primer oligonucleotides at 0.5–1.0 μM were prepared in 20 mM Na^+ or K^+ , unless otherwise indicated, in 10 μl . Samples were boiled for 5 min and then incubated for 1 h at 15 °C. Telomerase assay mixes were prepared on ice in 40 μl containing 10 μl 4 $\times$ buffer (200 mM Tris-HCl, pH 7.5, 40 mM MgCl_2), NaCl and KCl as indicated, 1 μl [$\alpha\text{-}^{32}\text{P}$]TTP (3,000 Ci mmol^{-1} , 10 mCi ml^{-1} ; NEN), 1 μl 25 μM TTP, 1 μl 5 mM dGTP, and 10 μl macronuclear extract. Preincubated oligonucleotide (1 μl) was added and the reaction mixture incubated at 15 °C for 30 min. Reactions were stopped by addition of 100 μl 10 mM Tris-HCl, pH 8.0, 25 mM EDTA, 2 μl 25 mg ml^{-1} yeast transfer RNA, 15 μl 5M ammonium acetate and 0.5 ml 100% ethanol. Gel electrophoresis and autoradiography were as described¹².

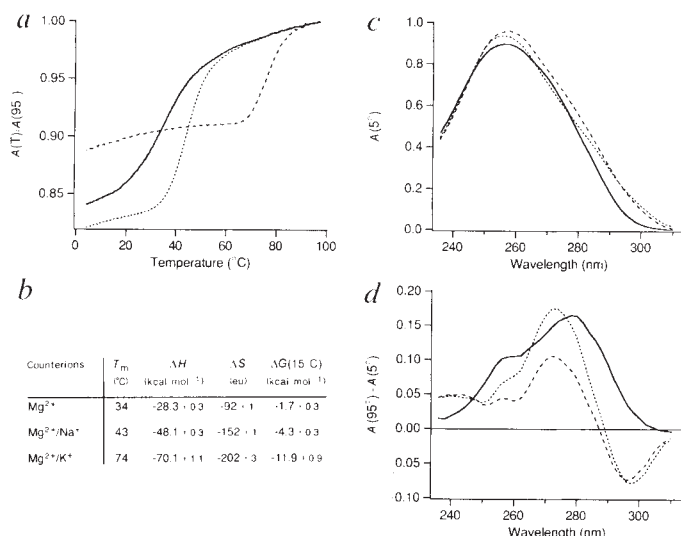
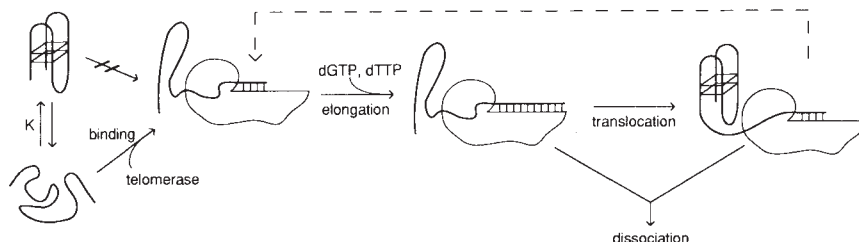


FIG. 3 Thermal denaturation and spectral analysis of oxy-4 in the presence and absence of Na⁺ or K⁺. *a*, Thermal denaturation of oxy-4 was monitored by the absorbance (*A*) at 270 nm in buffer containing 10 mM Mg²⁺ (solid

line), 10 mM Mg²⁺ plus 20 mM Na⁺ (dotted line), and 10 mM Mg²⁺ plus 20 mM K⁺ (dashed line). Absorbance was normalized to the absorbance at 95 °C. *b*, Thermodynamic parameters ΔH and ΔS were obtained by least-squares fitting of two independent data determinations, such as that shown in *a*, to a two-state transition with sloping initial and final baselines. The Mg²⁺/Na⁺ melting profile was independent of DNA concentration. The low-temperature profile for the Mg²⁺/K⁺ form was dependent on DNA concentration over an 8-fold range (0.55–4.5 μM). This is consistent with the observed presence of intermolecular species on native gels under similar conditions. For Mg²⁺/K⁺, only the concentration-independent transition above 40 °C was used for the data fitting. The reported errors reflect the uncertainty in the parameters based on an absolute absorbance error of ± 0.001 , and do not reflect any systematic errors that may be rather larger. *c*, Absorption spectra of oxy-4 at 5 °C, normalized to their absorbance at 95 °C. The absorption spectra at 95 °C were identical for the three forms, indicating that denaturation was complete at that temperature. The counterions are indicated as in *a*. *d*, Difference absorbance spectra for oxy-4 in the presence of the same counterions as in *a*. Thermal denaturation profiles and absorbance spectra were determined using a Gilford model 2400 spectrophotometer with a 4-cell sample changer, interfaced for data collection to an IBM personal computer. The buffer was 5 mM Mg²⁺-MOPS, pH 7.8. 1 mM Mg²⁺-MOPS stock solution was prepared by stirring MOPS free acid over magnesium oxide overnight and filtering. Samples were degassed before thermal denaturation in a Speed-Vac concentrator for 1 min, transferred to quartz microcells, and then overlaid with paraffin oil to minimize evaporation. The absolute absorbance of all three samples was ~ 1.0 at 270 nm, which corresponds to an oligonucleotide concentration of $\sim 3 \mu\text{M}$.

FIG. 4 Model for the effect of primer structure on *Oxytricha* telomerase. The most stable folded form (represented as a G-quartet structure⁸) is unable to bind to the RNA template of telomerase^{5,6}. The folded primer is in equilibrium with the unfolded form, which is able to bind the telomerase and is elongated by the addition of dG and T residues. Once the end of the internal RNA template is reached, a translocation step must occur before the next repeat is synthesized^{5,6}. If the primer/product becomes sufficiently long (at least 5 repeats for the G-quartet model), the portion of it that is not base-paired to the telomerase RNA template could undergo stable folding. The rate of dissociation of the primer/product from telomerase may be affected by the ability of the primer/product to form



such structures. Thus formation of folded structures might contribute to the regulation of telomere length *in vivo*.

is reduced because of the pre-equilibrium with the folded form. It is likely that the most stable folded form does not serve as a primer because binding to telomerase requires formation of Watson-Crick base pairs with the enzyme's internal RNA template^{5,6,11,13}. The more stable the folded form, the more the pre-equilibrium is shifted to the folded form, and the lower the apparent primer activity. This is consistent with four-repeat molecules being poor primers compared with two-repeat molecules, and with K⁺ reducing the activity with two-repeat primers even further. The absence of a large decrease in activity on addition of Na⁺ may mean that both the Mg²⁺ and the Mg²⁺/Na⁺ forms of the primer have an unfolded DNA concentration greater than the K_m for primer. But we cannot discount the possibility that the detailed folded structure formed in the presence of Na⁺ or Mg²⁺ can be opened by telomerase, whereas that in K⁺ cannot.

Formation of a folded structure in the portion of the primer not base-paired to the RNA template might also affect dissociation rates. G-quartet formation by a sufficiently long primer that emerges from the active site could alter the rates of any of the steps in the mechanism. Some evidence for such an effect can be seen in Fig. 2*a*. In the lanes for input primer oxy-1, 2 or 3, there is a discontinuity in the intensities of the products at a length that corresponds to oxy-5. When one repeat is base-paired to the internal template, the four unpaired repeats should still be able to fold into a structure. If G-quartet formation enhances dissociation of the one bound repeat, the five repeat products should be overrepresented relative to the four repeat products, as we observed.

Our results suggest that G-quartets may contribute to the regulation of telomere length *in vivo*. This regulation could be achieved in either of two ways: by direct inhibition of telomerase binding to long telomeric sequences, or by altering the dissociation rate of telomeric primers from telomerase. All known telomeric sequences are G-rich on the strand that is extended by telomerase, suggesting that the capacity of these sequences to adopt G-quartet structures is not a coincidence. □

Received 29 November 1990; accepted 4 March 1991.

- Blackburn, E. H. & Szostak, J. W. *A. Rev. Biochem.* **53**, 163–194 (1984).
- Zakian, V. A. *A. Rev. Genet.* **23**, 579–604 (1989).
- Greider, C. W. & Blackburn, E. H. *Cell* **43**, 405–413 (1985).
- Greider, C. W. & Blackburn, E. H. *Cell* **51**, 887–898 (1987).
- Greider, C. W. & Blackburn, E. H. *Nature* **337**, 331–337 (1989).
- Shippen-Lentz, D. & Blackburn, E. H. *Science* **247**, 546–552 (1990).
- Henderson, E., Hardin, C. C., Wolk, S. K., Tinoco, I. Jr & Blackburn, E. H. *Cell* **51**, 899–908 (1987).
- Williamson, J. R., Raghuraman, M. K. & Cech, T. R. *Cell* **59**, 871–880 (1989).
- Sundquist, W. I. & Klug, A. *Nature* **342**, 825–829 (1989).
- Sen, D. & Gilbert, W. *Nature* **344**, 410–414 (1990).
- Yu, G.-L., Bradley, J. D., Attardi, L. D. & Blackburn, E. H. *Nature* **344**, 126–132 (1990).
- Zahler, A. M. & Prescott, D. M. *Nucleic Acids Res.* **16**, 6953–6972 (1988).
- Shippen-Lentz, D. & Blackburn, E. H. *Molec. cell. Biol.* **9**, 2761–2764 (1989).
- Morin, G. B. *Cell* **59**, 521–529 (1989).
- Klobutcher, L. A. & Prescott, D. M. in *The Molecular Biology of Ciliated Protozoa* (ed. Gall, J. G.) 111–154 (Academic, Orlando, 1986).
- Gellert, M., Lipsett, M. N. & Davies, D. R. *Proc. natn. Acad. Sci. U.S.A.* **48**, 2013–2018 (1962).
- Oka, Y. & Thomas, C. A. Jr *Nucleic Acids Res.* **15**, 8877–8898 (1987).
- Zahler, A. M. thesis, Univ. Colorado (1990).
- Fersht, A. *Enzyme Structure and Mechanism* 2nd edn, 112 (Freeman, New York, 1985).
- Henderson, E., Moore, M. & Malcolm, B. A. *Biochemistry* **29**, 732–737 (1990).
- Raghuraman, M. K. & Cech, T. R. *Nucleic Acids Res.* **18**, 4543–4552 (1990).

ACKNOWLEDGEMENTS. We thank M. K. Raghuraman for discussions. J.R.W. was a fellow of the Jane Coffin Childs Memorial Fund for Medical Research T.R.C. is Investigator, Howard Hughes Medical Institute and American Cancer Society Professor. This work was funded by grants from the NIH (D.M.P. and T.R.C.).